Improved methods to recover HIV-1 integrated proviruses and integration sites

by

Kayla Eileen Delaney

Thesis presented in fulfilment of the requirements for the degree of Master of Science in Medical Virology from the Faculty of Medicine and Health Sciences at Stellenbosch University

Supervisor: Professor Gert Uves van Zyl

December 2020

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Full name : Kayla Eileen Delaney
Date : December 2020
Signature :

Abstract

Background

Long-lived HIV-1 infected cells that form part of the latent HIV-1 reservoir represent a major barrier to HIV-1 cure despite antiretroviral therapy (ART). The majority of these cells contribute to a persistent HIV-1 infection through clonal expansion. Many methods exist to study clonal expansion by identifying identical integration sites in different cells. However, the majority of proviruses are defective and until recently, methods did not exist to enable the simultaneous characterisation of integration sites and integrated proviruses, to identify which HIV-1 infected cell clones harbour intact proviruses. As recent methods are laborious and expensive, more efficient methods of linking intact proviral sequences and their cognate integration site are required. In addition, third-generation sequencing platforms allow for real-time long read-length sequence data which is ideal to characterise proviruses. These methods provide faster and simpler genome assembly than short read length second-generation sequencing platforms. Therefore, the overall aim of this study was to contribute to the HIV cure agenda, by improving our understanding of true HIV reservoirs by developing methods that would improve the characterisation of HIV-1 infected cell clones that harbour intact HIV genomes.

Methods

Intact proviral genomes are rare and attempts to enrich and link these proviral sequences to their integration site were investigated. The Integration Site Loop Amplification (ISLA) assay published by Wagner *et al*. (2014) was adapted for HIV-1 subtype C. Verification by Sanger sequencing showed that no integration sites were recovered and the “Premium whole genome amplification” method from Oxford Nanopore Technologies’ (ONT) third-generation sequencing technology was attempted as an alternative method. To investigate the utility of ONT sequencing, the “Amplicons by Ligation” method was utilised to sequence 9 near-full-length provisionally intact HIV-1 proviral sequences, previously sequenced by Illumina MiSeq. The consensus sequences for ONT sequencing were generated through a custom bioinformatic pipeline and compared to the Illumina MiSeq consensus sequences.

Results

The modified ISLA approach for HIV-1 subtype C or Premium whole genome amplification method did not succeed in recovering the HIV-1 proviral integration sites, of rare intact HIV-1 genomes. For near-full-length HIV genome sequencing, the “Amplicons by Ligation” protocol ONT sequencing achieved high coverage across the HIV genome and apart from hypervariable HIV-1 *envelope* there was a near perfect concordance between the consensus sequences generated with ONT and the previous Illumina MiSeq sequences, with an overall concordance of >99% for 8 out of 9 samples.

Conclusion

HIV-1 integration sites were not recovered in this study and efficient methods for simultaneous and efficient identification of intact proviral genomes and their integration sites remain unavailable. ONT sequencing allows for efficient and accurate sequencing of long fragments in real-time which may overcome technical barriers and eliminate laborious, time-consuming and expensive methods currently used for integration site identification and near-full-length HIV-1 genome sequencing. As part of the research towards future possible HIV cures, it remains a priority to investigate the persistence of cells harbouring intact HIV-1 genomes, and the role of clonal cellular proliferation in their survival.

Opsomming

Agtergrond

Die grootste struikelblok om MIV-1 te genees, ten spyte van antiretrovirale terapie (ART), is langlewende MIV-1 geïnfekteerde selle wat deel uitmaak van die latente MIV-1 reservoir. Die meerderheid van hierdie selle dra by tot die handhawing van MIV-1 infeksie deur klonale vermeerdering. Verskeie bestaande metodes word gebruik om klonale vermeerdering te ondersoek deur die identifisering van identiese integrasie lokusse in verskillende selle. Nietemin is die meerderheid van proviruses defektief. Tot onlangs, het metodes wat die gesamentlike karakterisering van integrasie lokusse en geïntegreerde provirusse, moontlik maak, om vas te stel watter MIV-1 geïnfekteerde selklone intakte provirusse huisves, nie bestaan nie. Aangesien onlangse metodes baie moeisaam en duur is, is meer doeltreffende metodes wat intakte provirusse en hul onderskeie integrasie lokusse met mekaar verbind, benodig. Boonop laat derde generasie volgordebepalingstegnologie blitsvinnige lang-leeslengte volgordebepaling toe, wat ideal is vir die karaktisering van provirusse. Hierdie metodes laat vinniger en makliker genoom samestelling toe in teenstelling met kort-leeslengte tweede generasie volgordebepalingstegnologie. Gevolglik is die oorhoofse doel van hierdie studie om tot die MIV genesing-agenda by te dra deur ons begrip van MIV-1 reservoirs te verbeter, met die ontwikkeling van metodes wat die karakterisering van MIV-1 geïnfekteerde selklone wat ongeskonde MIV genome huisves, laat vooruitgaan.

Metodes

Ongeskonde provirale genome is skaars en pogings om hierdie provirusse te verryk en te verbind met hul onderskeie integrasie lokusse is ondersoek. Die “Integration Site Loop Amplification” (ISLA) metode wat deur Wagner *et al*. (2014) gepubliseer is, is aangepas vir MIV-1 subtipe C. Die nagaan hiervan deur Sanger volgordebepaling dui daarop dat geen integrasie lokusse herwin is nie; en is die “Premium whole genome amplification” metode van Oxford Nanopore Technologies (ONT) derde-generasie volgordebepaling as alternatiewe metode gebruik. Om die gebruiklikheid van ONT volgordebepaling te ondersoek, was die “Amplicons by Ligation” metode gebruik om 9 naby-vollengte voorlopig-ongeskonde MIV-1 provirusse, waarvan die volgorde voorheen deur Illumina MiSeq bepaal is, gebruik. Die konsensusvolgordebepaling vir ONT is gegenereer deur ‘n pasgemaakte bio-informatika pyplyn en vergelyk met die Illumina MiSeq konsensus volgordebepalings.

Bevindinge

Die aangepasde ISLA MIV-1 subtipe C of “Premium whole genome amplification” metodes was nie suksesvol met die herwinning van die provirale integrasie lokusse van raar ongeskonde MIV-1 genome nie. Die “Amplicons by Ligation” metode van ONT was gebruik om die volgorde van naby-vollengte MIV genome te bepaal, met ‘n hoë leesdekking oor die MIV genoom. Buiten vir die hoogsveranderlike MIV-1 *membraan* geen was daar ‘n amper-perfekte ooreenstemming tussen die konsensus volgordebepaling gegenereer met ONT en die Illumina MiSeq volgordebepaling, met ‘n totale ooreenkoms van >99% vir 8 van die 9 monsters.

Samevatting

MIV-1 integrasie lokusse was nie ingesluit in hierdie studie bevindings nie en doeltreffende metodes vir gesamentlike identifisering van ongeskonde provirale genome en hul integrasie lokusse bly onverkrygbaar. ONT volgordebepaling laat doeltreffende en akkurate volgordebepaling en lang-leeslengte blitsvinnig toe. Dit kan moontlik tegniese versperrings van huidige tydrowende en duur metodes, om integrasie lokusse en naby-vollengte MIV-1 genoom volgordes te bepaal, oorkom. As deel van die navorsing oor moontlike genesings vir MIV, bly dit ‘n prioriteit om die voortlewing van selle wat ongeskonde MIV-1 genome huisves, en die rol van klonale sellulêre vermeerdering in hul oorlewing te ondersoek.

Acknowledgements

I would like to acknowledge the following individuals and organisations for their support and contributions towards the completion of this degree.

a) Professor Gert Uves van Zyl for his guidance, support, mentorship and optimism throughout a challenging Master's project. I will be forever grateful for his contribution to this degree and for providing numerous opportunities to develop my technical skills and strategic thinking capabilities as a researcher.

b) Professor Susan Engelbrecht for her assistance with Nanopore Sequencing training, her invaluable contribution in the form of lab experience which assisted in assay optimisation and for inviting me to be part of the NGS-SA COVID-19 research team.

c) The current and previous members of the research team:
   - Shahieda Isaacs, Kirsten Veldsman and Mary Grace Katusiime for their encouragement, assistance with assay training and sharing of personal lab experiences.
   - Carli van Zyl for her assistance with assay training and for her overall support, encouragement and optimism as a fellow MSc student and a great friend.
   - A massive thank you to Ian Botha for always assisting with problem solving, providing guidance and expertise with assay optimisation.

   I am sincerely grateful for each team members support, encouragement and contribution to skills development, this project would not have been possible without them.

d) All the students and staff in the Division of Medical Virology. And a special thank you to Bronwyn Kleinhans for taking time out of her PhD project to assist with training and sharing valuable knowledge.

e) To the following organisations for funding, the Harry Crossley Foundation and Poliomyelitis Research Foundation. I am sincerely grateful for their contribution as it facilitated the development of my skills and also to the advancement of research.

f) My family and friends, especially my biggest supporters and cheerleaders, my parents, David Delaney and Yvette Delaney. Thank you for your continuous support, encouragement and love throughout this postgraduate journey.

g) To our Heavenly Father, whose love endures and provided me with the strength, perseverance and optimism to continue especially during the challenging times.

Research Outputs

List of research outputs from the work completed during this Master's project.

Conference attendance:

Poster presentation:

Delaney, K.E., Wright, I., Katusiime, M.G., Engelbrecht, S., van Zyl, G.U. 2020. High Coverage Oxford Nanopore Technology Sequencing Provides Accurate Near-Full-Length Consensus Sequences. *Virology Africa 2020*, Cape Town, South Africa.

List of research outputs from the work completed on the South African COVID-19 epidemic in the Western Cape.

Poster presentation:

Engelbrecht, S., Delaney, K.E., Kleinhans, B., Wilkinson, E., Tegally, H., Stander, T., van Zyl, G.U., Preiser, W., de Oliveira, T. 2020. Molecular epidemiology and genetic diversity of SARS-CoV-2 in Cape Town, South Africa. *Stellenbosch University 64th Annual Academic Year Day*, Cape Town, South Africa.

A brief report on the findings of this study is provided in Addendum A.

Table of contents

Improved methods to recover HIV-1 integrated proviruses and integration sites

List of Abbreviations

AIDS	Acquired Immunodeficiency syndrome
APOBEC3G	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G
ART	Antiretroviral therapy
ASIC	Application-specific integrated circuit
blastn	Nucleotide blast on NCBI platform
$CaCl_2$	Calcium Chloride
CAF	Central analytical facility
cART	Combination antiretroviral therapy
CCR5	C-C chemokine receptor type 5
CCS	Circular continuous sequence
CD4	Cluster of differentiation 4 (glycoprotein)
CHER	Children with HIV early antiretroviral therapy
CLR	Continuous long read
CRF	Circulating recombinant form
CXCR4	C-X-C chemokine receptor type 4
ddNTPs	Dideoxynucleotide triphosphates
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleoside triphosphates
DMSO	Dimethyl sulfoxide
EB	Elution buffer
EDTA	Ethylenediaminetetraacetic acid
FB	Flush buffer
FBS	Fetal bovine serum
FLT	Flush Tether
GuHCl	Guanidinium hydrochloride
GuSCN	Guanidinium isothiocyanate

HIV	Human Immunodeficiency virus
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ISLA	Integration site loop amplification
kb	Kilobase pairs
LFB	Long fragment buffer
LN_2	Liquid nitrogen
LNB	Ligation buffer
LTR	Long terminal repeat
MCS	Multiple cloning site
MDA	Multiple displacement amplification
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
NEB	New England Biolabs
Nef	Negative factor
NFL	Near-full-length
NRTI	Nucleoside or nucleotide reverse transcriptase inhibitor
NNRTI	Non-nucleoside reverse transcriptase inhibitor
ONT	Oxford Nanopore Technologies
p6-PR-RT	Amplification of HIV-1 from the p6 region to reverse transcriptase
PacBio®	Pacific Biosciences
PBMC	Peripheral blood mononuclear cell
PCR	Polymerase chain reaction
PIs	Protease inhibitors
PIC	Preintegration complex
ProK	Proteinase K
QC	Quality control
RCF	Relative centrifugal force

RNA	Ribonucleic acid
RPM	Revolutions per minute
RPMI	Roswell Park Memorial Institute (media)
SFB	Short fragment buffer
SGA	Single genome amplification
SMRT	Single-molecule Real-time
SNP	Single nucleotide polymorphisms
S.O.C.	Super Optimal broth with Catabolite repression (media)
SQB	Sequencing buffer
T_m	Melting temperature
TE	Tris and EDTA buffer
Tris-HCl	Trisaminomethane hydrochloride
UNAIDS	Joint United Nations Programmed on HIV/AIDS
Vif	Viral infectivity factor
VOA	Viral outgrowth assay
Vpr	Viral protein r
Vpu	Viral protein u
WHO	World Health Organisation
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
ZMW	Zero-mode waveguide

Table of Figures

Table of Tables

Chapter 1

1 Introduction

1.1 HIV burden globally and in South Africa

The Joint United Nations Programmed on HIV/AIDS (UNAIDS) devised the 90-90-90 strategy which aimed to end the AIDS epidemic, caused by HIV infection, by the year 2030. By 2020, the strategy aimed to achieve the following; 90% of the individuals living with HIV should know their status, of those, 90% should be receiving sustained antiretroviral therapy (ART) and 90% of people receiving ART should be virally suppressed. In 2019, the UNAIDS global statistics reported 38 million individuals were estimated to be living with HIV of which approximately 81% knew their status, approximately 67% were accessing ART and 59% were virally suppressed (UNAIDS, 2020). South Africa's National Department of Health adopted the 90-90-90 strategy in December 2014 as South Africa is reported to have the largest HIV epidemic worldwide and AIDS is one of the leading causes of death in the country (Health Systems Trust, 2016). The UNAIDs reported that approximately 7.7 million individuals were living with an HIV infection by 2018 in South Africa which accounted for almost 21% of the HIV infections globally that year. The latest 90-90-90 statistics for South Africa were released in 2018 and show that 90% of individuals knew their status, 62% of adults and 63% of children were receiving ART and 54% were virally suppressed (UNAIDS, 2020). Steady progress towards achieving the 90-90-90 strategy goal has been made, however, it remains to be known whether these goals have been achieved by the end of this year. The development of a cure for HIV remains a priority research area as treatment with ART alone is not sufficient at eradicating HIV infection. Current cure research is aimed at identifying strategies that allow for effective drug-free control of HIV infection as a 'functional cure' and identifying drugs that can eradicate HIV as a 'sterilising cure'. Vaccine development and drug resistance research also remain important areas of focus. The information generated in different disciplines aim to ultimately contribute towards ending the epidemic.

1.2 HIV classification, structure and replication

1.2.1. HIV classification

HIV is classified in the *Lentivirus* genus within the family of Retroviridae. Based on the genetic characteristics and differences present in the viral antigens, two distinct HIV species have been identified, namely HIV-1 and HIV-2. HIV-1 is responsible for the global pandemic

and contributes to 95% of the HIV infections worldwide. HIV-2 has a very low prevalence outside of West Africa, a much lower rate of transmission and a slower disease progression than HIV-1 (Zheng *et al*., 2004).

HIV-1 is classified into four highly divergent groups: M (major), O (outlier), N (non-M, non-O) and P. Group M is the pandemic branch that is further divided into subtypes based on the phylogenetic associations of the sequences, namely; A1, A2, A3, A4, A6, B, C, D, F1, F2, G, H, J, K and also includes circulating recombinant forms (CRFs). Unique recombinant forms occur as a result of recombination events between known and recognised subtypes and are constantly emerging and result in the formation of CRFs of HIV-1 once transmitted to three or more individuals who are not epidemiologically related (Leitner *et al*., 2005, Olabode *et al*., 2019). Currently, 102 CRFs are represented on the Los Alamos National Lab HIV Sequence Database [www.hiv.lanl.gov]. HIV-1 subtype C is the most prevalent subtype and accounts for approximately 50% of the infections worldwide (Taylor *et al*., 2008, Günthard & Scherrer, 2016, Gartner *et al*., 2020) and is predominant in Southern and Eastern Africa, India and Ethiopia (Gartner *et al*., 2020).

1.2.2. HIV-1 structure

Mature HIV virions, as depicted in Figure 1.1B, measure approximately 100-120 nm in diameter and are surrounded by an outer lipid membrane which serves as its envelope. The 9.2 - 9.6 kilobase pair (kb) HIV genome (German Advisory Committee Blood, 2016) consists of two identical copies of positive single-stranded RNA which encodes for 16 proteins that can be divided into three protein categories; structural, regulatory and accessory. Structural proteins are essential in manufacturing components of the retroviral particle and are products of *env*, *gag* and *pol* genes depicted in Figure 1.1A. The outer surface of the virion is covered with 72 spikes that are produced by trimers of two *env* glycoproteins. The spikes consist of the surface gp120 trimer which is anchored to the membrane by the gp41 transmembrane protein. *Gag* produces internal structural proteins p17, p24 and p7 which are responsible for producing the matrix, capsid and nucleocapsid, respectively. The cone-shaped capsid of a mature virion houses the viral genome and the encoded viral enzymes; protease, reverse transcriptase and integrase which play essential roles in viral replication and are produced by the *pol* gene (Engelman & Cherepanov, 2013, Li & De Clercq, 2016). Regulatory proteins are encoded by *tat* (trans-activator of transcription) and *rev* (regulator of viral expression) and are essential in HIV replication initiation. Accessory or auxiliary

proteins; Nef (negative factor); Vpr (viral protein r), Vpu (viral protein u) and Vif (viral infectivity factor) impact on viral replication, virus budding and the pathogenesis of HIV. The HIV genome is flanked on both ends by long terminal repeat (LTR) sequences which collectively include the unique 5' (U5), repeat (R) and unique 3' (U3) regions. The 5'LTR region encodes for the promoter for the transcription of viral genes (German Advisory Committee Blood, 2016).

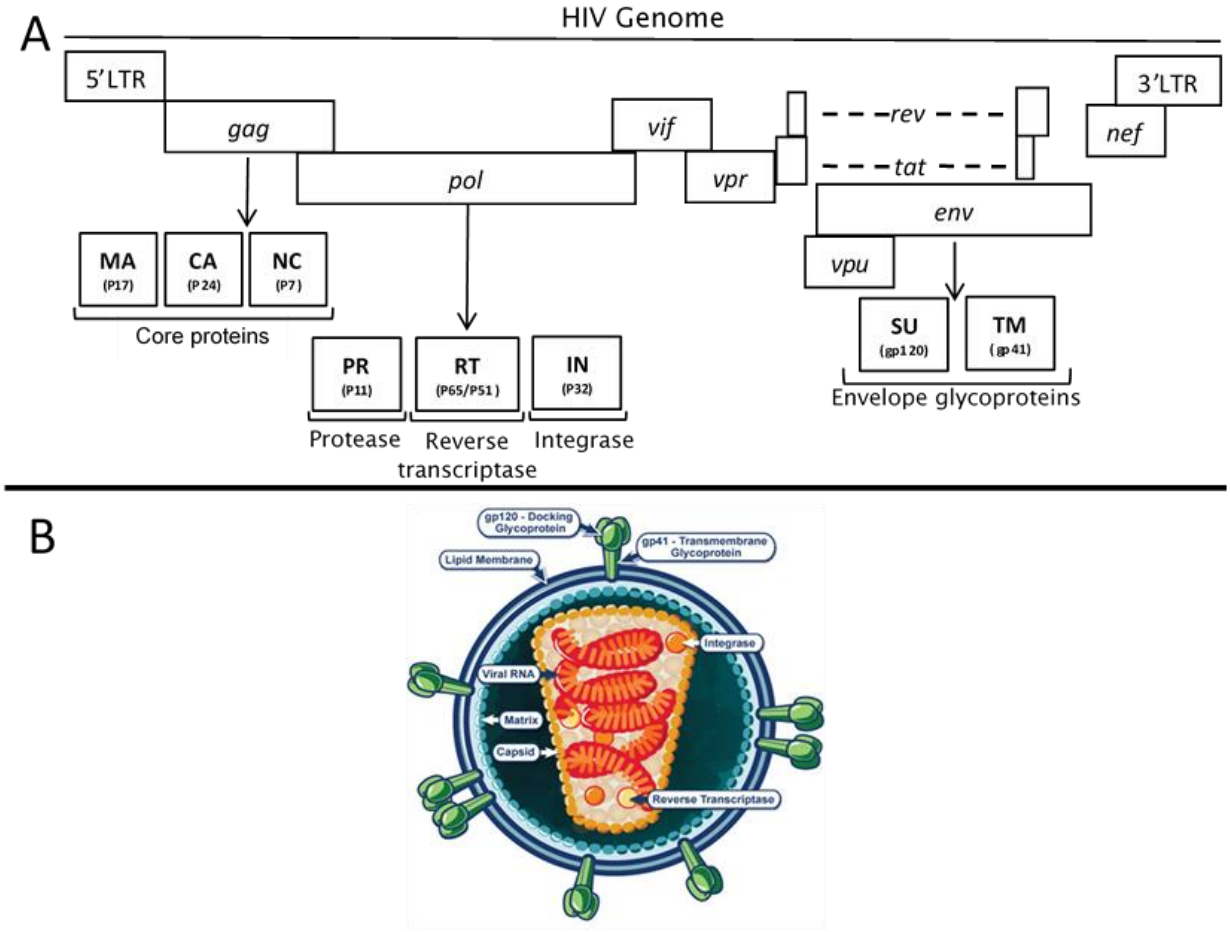

Figure 1.1: The genome and structure of a mature HIV-1 virion. (A) Proviral HIV-1 genomic map showing the viral genes and the resulting structural and regulatory viral proteins. (B) A mature HIV-1 virion. (Adapted from sources: Nkeze *et al*., *2015* and National Institute of Allergy and Infectious Diseases, National Institutes of Health, 2009).

1.2.3. HIV-1 replication

HIV primarily infects CD4+ T cells but is also capable of infecting other immune cells that have the CD4 receptor which includes monocytes, macrophages and dendritic cells (Engelman & Cherepanov, 2013). Viral replication occurs through a series of steps that begins with the productive engagement of the virus with host cell surface receptors and ends when nascent viral particles mature into infectious virions as shown in Figure 1.2. The viral

entry or fusion begins with the attachment of the host cell CD4 receptor via the C4 domain of gp120 which evokes a conformational change in CD4 and allows gp120 to bind to co-receptor CCR5 or CXCR4. The gp41 trimer mediates the fusion between the HIV virion and the host immune cell. Reverse transcription is initiated and the single-stranded RNA is converted into double-stranded DNA by the reverse transcriptase enzyme. The double-stranded DNA and integrase form part of the preintegration complex (PIC) which enters the host's cell nucleus. Integrase catalyses the integration of the viral DNA into the host cell's genomic DNA. Successful integration relies on the occurrence of three sequence specific events which are facilitated by integrase; viral DNA is assembled, the 3' ends of the viral DNA are endonucleolytically processed and the viral and host cellular DNA are joined. The integration step is critical as it finalises the infection of the host cell and the establishment of a persistent HIV infection. The integrated HIV genomes are referred to as proviruses. Viral gene expression follows where host cellular machinery such as transcription factors are utilised to transcribe new viral messenger RNA (mRNA) from the integrated proviral DNA. The transcribed mRNA contains the necessary information to produce new viral proteins that are required to assemble a new HIV virion. The viral RNA exits the cell nucleus and is translated into virus proteins which assemble at the inner plasma membrane of the host cell to allow for budding off and the formation of an immature HIV virion which is non-infectious. The newly formed immature HIV virion is released and the viral enzyme protease initiates proteolysis which cleaves the polyproteins to create a mature infectious virion (Arts & Hazuda, 2012).

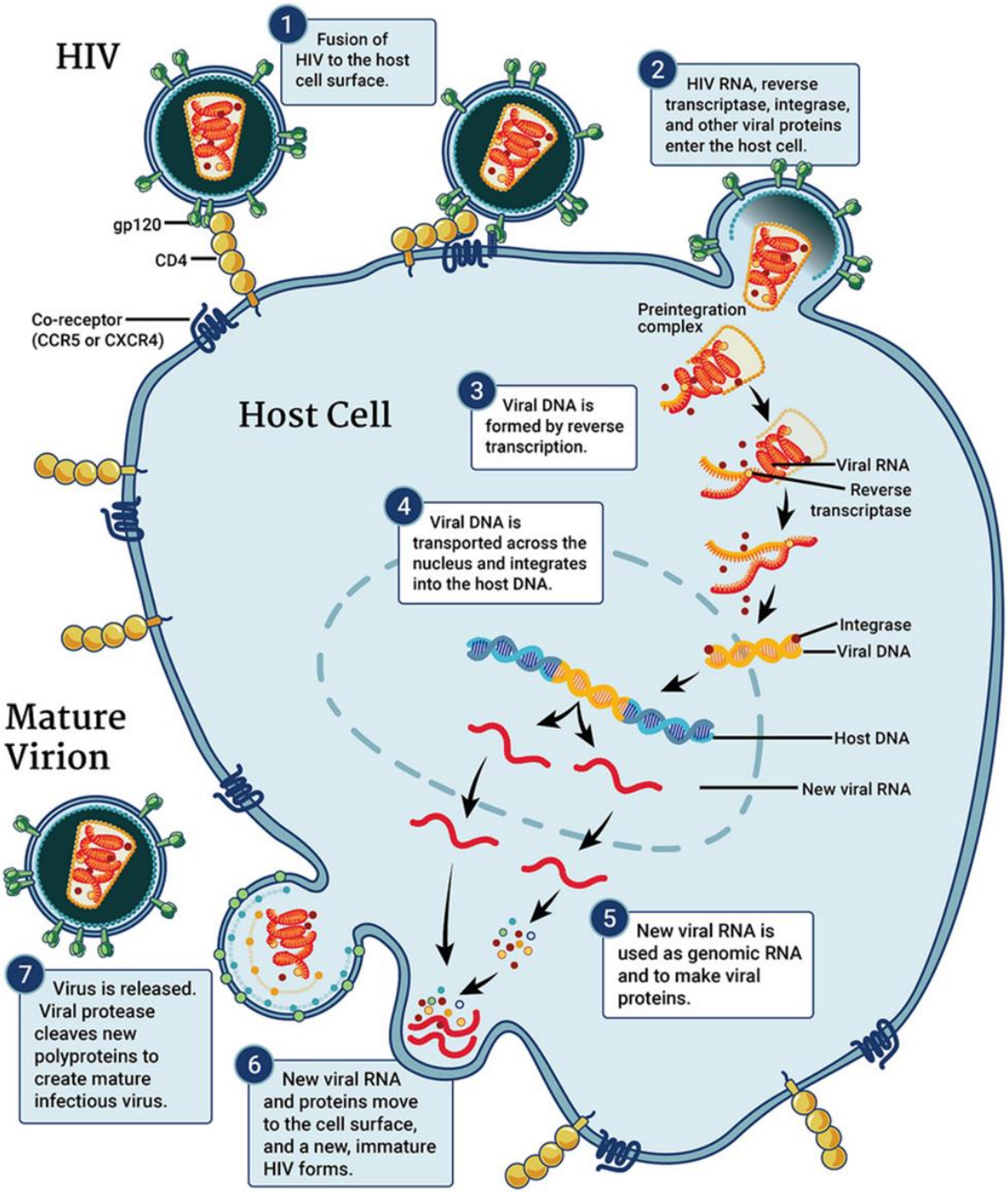

Figure 1.2: HIV-1 replication cycle (Source: National Institute of Allergy and Infectious Diseases, 2018. [Online]: https://www.niaid.nih.gov/diseases-conditions/hiv-replication-cycle).

1.3. HIV-1 infection and immune function

HIV primarily targets activated CD4+ T cells as this leads to a productive viral infection. These cells are involved in the co-stimulation and regulation of humoral and cellular immunity however, viral infection of these cells disables these functions. The HIV-infected CD4+ T cells are destroyed by the cytotoxic CD8+ T cells of the immune system. HIV

infection consists of several stages and without the intervention of ART can progress to AIDS in a median of ten years as indicated in Figure 1.3 (Tobin & Aldrovandi, 2013). The acute stage of infection is characterised by a very high HIV-1 RNA plasma viral load and a low CD4+ T cell count as the infected CD4+ T cells are destroyed by the immune system (Okoye & Picker, 2013). Regeneration from CD4+ memory T cells of the immune system restores the CD4+ T cell numbers and functions above the threshold that is associated with immunodeficiency, viral replication is controlled and the HIV-1 RNA plasma viral load decreases by 100 to 1 000-fold. The HIV-1 RNA plasma viral load and CD4+ T cell count remain relatively stable at this viral set-point for years (Tobin & Aldrovandi, 2013). Over time, immune exhaustion occurs and homeostasis of the CD4+ T cell population fails (Okoye & Picker, 2013) and the number of cells decline to below the level that is required to prevent opportunistic infections and the viral load increases which is concurrent with the onset of HIV related symptoms which leads to AIDS and death.

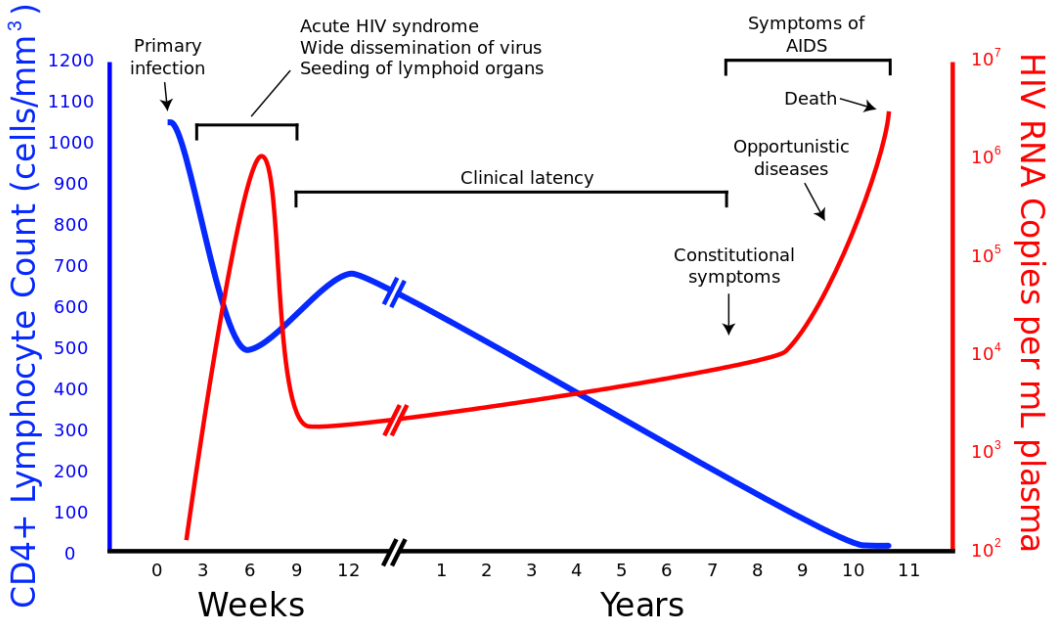

Figure 1.3: Progression of HIV-1 infection measured by CD4+ T cell count and HIV-1 RNA plasma viral load in an ART naïve individual. (Adapted from source: Kogan & Rappaport, 2011).

1.4. HIV-1 infection in paediatrics

The progression of HIV infection is significantly different between adults and children and could be related to the reported differences in their immune systems. In adults, without ART intervention, progression to AIDS or death usually occurs within ten years following primary infection as discussed above. In contrast, without the intervention of ART, children living

with HIV have shown higher rates of mortality and faster disease progression as shown in Figure 1.4 (Tobin & Aldrovandi, 2013, Martinez *et al*., 2015). Approximately 20 to 30% of HIV-infected children usually die within the first year of life (Martinez *et al*., 2015). Furthermore, HIV-1 RNA plasma viral loads greater than 100 000 copies per millilitre have been reported in perinatally infected infants months after infection and are typically 10-fold higher than has been reported in adults. In children who do not progress to AIDS or death, a viral set-point is only achieved at five years of age. A lower number of circulating target HIV cells, CD4+ T cells with the CCR5 co-receptor (Shalekoff *et al*., 2004) are present in children although a high proportion of these cells have recently been reported in the gut (Bunders e*t al*., 2012). Cellular immune responses to HIV infection also differ. In adults, prior to peak viremia, virus specific CD8+ T cells are present in blood and the cytotoxic T cell responses are associated with the observed decrease in viral load. In children, however, these cytotoxic responses are only detected after six months of age and viral replication is controlled to some extent by non-cytolytic CD8 cells (Martinez *et al*., 2015). Overall, the lack of viraemic control observed in children may be attributed to the immature immune system which is comprised of deficient HIV specific CD4+ T cell responses, ineffective CD8+ T cell responses and a delay in antibody-dependent cell-mediated cytotoxicity.

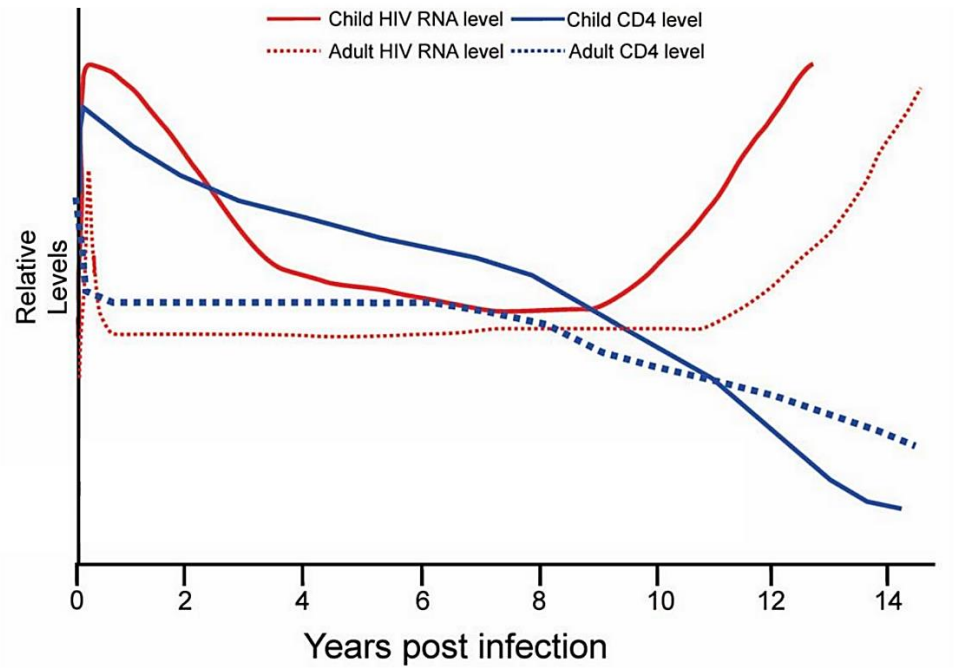

Figure 1.4: Comparison of HIV-1 disease progression in adults and children without the intervention of ART. The relative HIV-1 RNA plasma viral load is shown in red and the CD4+ T cell count is shown in blue. (Adapted from source: Tobin & Aldrovandi, 2013) *Image used with permission from Immunological Reviews.*

1.5. Treatment of HIV-1 infection

1.5.1. Antiretroviral drugs

ART inhibits HIV replication by targeting various steps that are specific to the viral replication cycle. Currently, six distinct classes of ART have been licenced for use in patients and are classified according to their molecular mechanism and include nucleoside analogue reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs), integrase inhibitors, protease inhibitors (PIs), fusion/entry inhibitors and chemokine coreceptor antagonists which consist of two subclasses, namely the CCR5 antagonist and the CXCR4 antagonist (Arts & Hazuda, 2012). Monotherapy was used as the initial treatment strategy for HIV infection however, sustained therapeutic success could not be achieved and failure was attributed to the build-up of antiretroviral drug resistance (Engelman & Cherepanov, 2013). The standard of care evolved into what is now known as combination antiretroviral therapy (cART) where at least three different antiretroviral drugs from two different classes are given to HIV infected individuals. The implementation of cART regimens improved the morbidity and mortality that was originally associated with HIV infection and AIDS. A rapid suppression of viral replication and a decrease in HIV-1 RNA plasma viral loads to below the limit of detection in sensitive clinical commercial assays (20 - 50 HIV-1 RNA copies/ml) were observed in individuals on suppressive cART (Vanhamel *et al*., 2019). cART is effective at preventing the infection of new cells, and when current regimens are prescribed it limits viral evolution and the generation of resistant escape variants, in the large majority of patients. Nevertheless, long-lived HIV reservoirs persist and therefore cART does not provide a curative strategy for HIV infection. Upon cART cessation HIV usually rebounds quickly from long-lived cellular viral reservoirs. Therefore, lifelong treatment for individuals with HIV infection is required and is the current standard of care.

1.5.2. Benefits of early treatment with cART

Multiple studies have reported significant benefits of early cART initiation in both adults and children. A study conducted in adults by Le and colleagues (2013) showed that treatment within the first four months of infection significantly correlated with a higher restoration of the CD4+ T cell counts in the blood to >900 cells/ml (median value of 992 CD4+ T cells/ml was measured for HIV seronegative individuals whereas with treatment initiation during chronic infection, enhanced restoration of these cells was less likely to occur. Therefore, cART initiation early in acute infection significantly improves the probability of immune recovery.

Approximately, 5 -15% of adults treated during acute infection maintained viraemic control following cART interruption whereas therapy initiation during chronic infection typically resulted in viral rebound following treatment interruption (Sáez-Cirión *et al*., 2013, Stohr *et al*., 2013). Another study reported that early initiation and lengthened cART adherence resulted in a slower disease progression and time to viral rebound (SPARTAC Trial Investigators *et al*., 2013). Similar findings have been observed in early cART treated children with HIV infection. The randomized children with HIV early antiretroviral (CHER) trial investigated the benefits of early treatment initiation and the effect of treatment interruption at two time points. Overall, the study reported that the risk of death was reduced by 75% with early initiation of cART and similar to SPARTAC Trial (2013), a longer duration on cART prior to treatment interruption resulted in slower time to viral rebound which could potentially assist in the development of a sterilising or functional cure (Cotton *et al*., 2013). Furthermore, Kuhn *et al*. (2018) reported higher levels of HIV-1 DNA in children who started cART at ≥4.5 months of age than those who started earlier and also provided strong evidence of an association between early cART initiation and a reduced size of the viral reservoir several years later in children on suppressive therapy. Taken together, initiation of cART shortly after infection allows for a greater recovery of CD4+ T cell numbers, assists in lowering virologic set points, limits viral diversity and the size of the reservoir of persisting HIV (Buzon *et al*., 2014) while preserving immune responses in children and adults with HIV.

1.6. HIV-1 reservoir

1.6.1. Establishment of the HIV-1 reservoir

Current cART regimens are successful in suppressing viral replication in the majority of HIV-infected individuals by targeting viral enzymes or viral entry. However, it cannot eradicate the virus as the long-lived HIV reservoir is established early in infection. Activated CD4+ T cells are the preferential targets for productive HIV infection, but they have short lifespans due to the cytopathic effects of the virus or cytotoxic T lymphocyte-mediated killing (Coffin & Swanstrom, 2013) as shown in Figure 1.5. The majority of newly HIV-infected CD4+ T cells die within the first 24 – 48 hours. The remaining cells either die within the next 2 – 4 weeks or revert to a resting state and differentiate into long-lived memory CD4+ T cells (Coffin & Swanstrom, 2013, Maldarelli *et al*., 2014) depicted in Figure 1.5. The main immune cells comprising the reservoir are, therefore, memory CD4+ T cells which have an estimated half-life of 43 – 44 months. Due to the stability of the reservoir cells, cART would have to be administered for >73 years in order for the latent reservoir to decay to zero (Siliciano *et al*.,

2003, Hosmane *et al*., 2017). The HIV reservoir, therefore, creates a major barrier to cure as the cells are stable, persist throughout the duration of suppressive cART and are capable of producing replication-competent virus upon cART cessation. The presence of a latent HIV reservoir is supported by clinical evidence showing that poor adherence or interruption of cART leads to viral rebound. Therapy is therefore administered for life in order to suppress viral replication and slow the progression of the disease. Recent studies have shown that various mechanisms may play a role in reservoir persistence: 1) intact proviruses are able to persist in long-lived latently infected cells, 2) possible low level and ongoing viral replication may occur due to insufficient levels of intracellular drug concentrations (Lorenzo-Redondo *et al*., 2016) although data provided by van Zyl and colleagues in 2017 provides strong evidence against ongoing viral replication and evolution and 3) the proliferation of HIV-1 infected cells also known as clonal expansion (Maldarelli *et al*., 2014, Wagner *et al*., 2014, Simonetti *et al*., 2016) which is currently a prominent area of ongoing research.

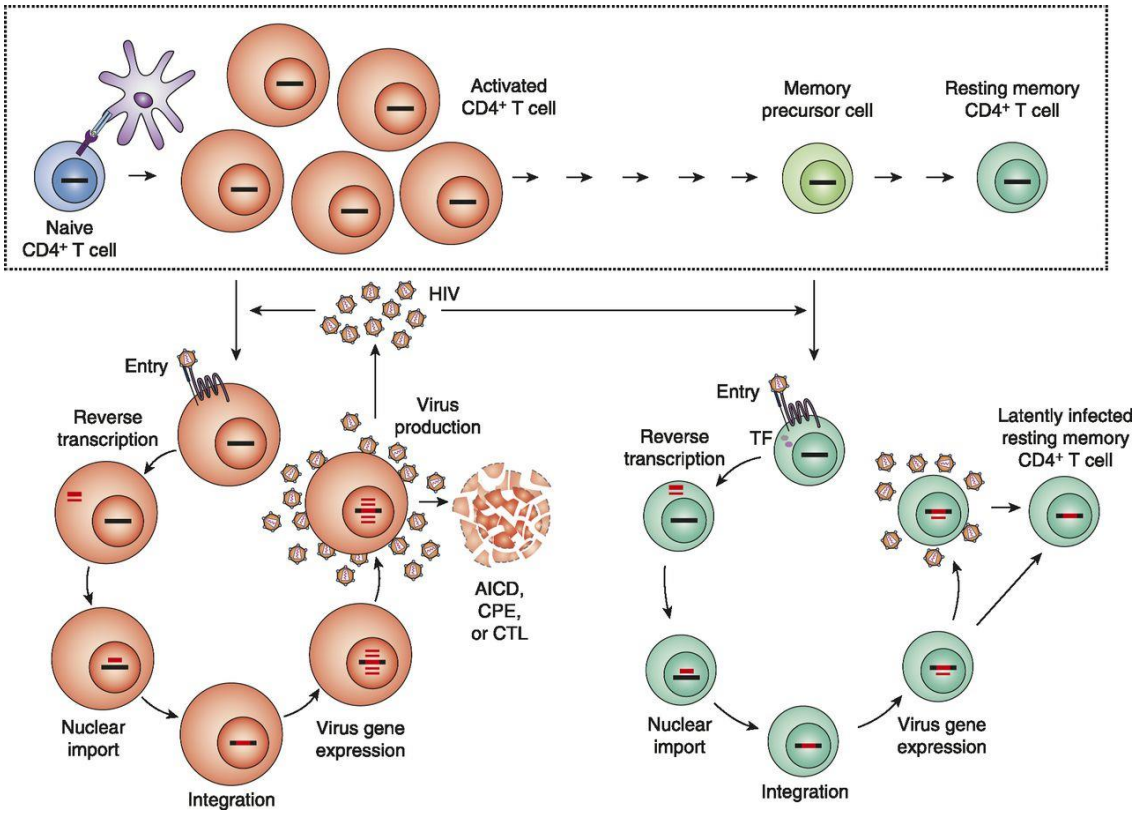

Figure 1.5: Comparison of a productive and latent HIV infection. The normal process of memory cell generation is shown in the boxed area. HIV infects activated CD4+ T cells and results in a productive HIV infection where most infected cells die. A proportion of HIV-infected activated CD4+ T cells can revert to resting state and form part of the latent reservoir where they contribute to a persistent HIV infection. (Source: Murray *et al*., 2016) *Image used with permission from The Journal of Immunology*.

1.6.1.2. The HIV-1 latent reservoir undergoes clonal expansion

Long-lived cells infected by HIV before the initiation of cART are the most important part of the latent HIV reservoir and are capable of persisting for years despite fully suppressive treatment. Evidence has shown that within a patient, a large number of infected CD4+ T cells have proviral integration into the exact same position in the human genome, which is extremely unlikely to have resulted from multiple integration events. This provides strong support for clonal expansion of HIV-infected cells. These expanded clones can be detected in patients on long-term therapy and may show integration in the same orientation as growth genes (Wagner *et al*., 2014 and Maldarelli e*t al*., 2014). HIV integration is a relatively nonspecific process; however, evidence has shown that integration preferentially occurs in expressed genes throughout the genome (Schröder *et al*., 2002; Han *et al*., 2004) which can be explained by the proliferation of infected cells after integration.

HIV integration can involve the insertion of an intact, full-length, replication-competent provirus or a defective provirus into the host genome prior to latency. However, the integration of a defective provirus is more likely to occur, as cells carrying defective proviruses are less likely to die from viral cytopathic effects or lysis by host effector cells. More than 90% of latent HIV proviruses are defective in cART suppressed individuals. Defective proviruses are characterised by the presence of deletions, insertions, inversions or mutations in the HIV genome that are introduced by the error-prone reverse transcriptase enzyme, template switching during reverse transcription, APOBEC3G-induced hypermutations or point mutations (Ho *et al*., 2013, Bruner *et al*., 2016). Due to the high prevalence of defective latent proviruses, it was thought that the majority of clonally expanded cells would harbour defective HIV proviruses. Both defective and intact proviruses survive through clonal expansion. However, it was previously not known whether intact proviruses were able to survive by clonal expansion without being purged by the immune system.

Simonetti *et al*. (2016) was the first to report that an HIV infected individual with a highly expanded clone carried an intact provirus that was responsible for persistent, infectious viremia. In addition to this, three independent studies have shown that approximately 56% of cells that harbour replication-competent HIV proviruses undergo clonal expansion (Bui *et al*., 2017, Lorenzi *et al*., 2016, Hosmane *et al*., 2017). The study carried out by Bui and colleagues in 2017 demonstrated that cells carrying intact proviruses can clonally

expand *in vivo* and recovered identical sequences in six out of eight participants. Lorenzi and colleagues (2016) showed that more than 50% of replication-competent isolates from four patients on suppressive cART had identical *env* sequences to other isolates from the same individual. However, this study did not investigate other regions in the HIV genome and although *env* showed identical sequences, the sequence may differ elsewhere in the genome (Laskey *et al*., 2016). Hosmane *et al*. (2017) not only showed that 57% of latently infected cells carried identical *env* sequences to other isolates from the same patient but further showed that these sequences were identical throughout the entire genome. Taken together, these studies provide strong support for clonal proliferation as a mechanism of survival of replication-competent proviruses and suggest that the majority of the latent reservoir is likely maintained by clonal proliferation.

1.6.1.2.1. Assays investigating clonal proliferation

Clonal expansion is reported as the major mechanism of latent reservoir persistence and supporting evidence has been provided by the identification of identical integration sites in infected cells using integration site assays. Considering the size of the human genome statistically, two infected cells harbouring identical proviral integration sites are more likely to have emerged as a result of a single infection event followed by the proliferation of the infected cell, as opposed to two separate integration events. Therefore, infected cells with identical HIV integration sites are indicative of clonal expansion.

Most assays used to determine the human genome integration position of particular proviruses include short stretches near the respective 5’ or 3’ ends of the HIV genome only and do not enable one to assess proviral intactness nor confidently link the integration sites and their respective proviruses. It is likely that the majority of the expanded clones detected in these assays represent defective proviruses (Cohn *et al*., 2015). Approaches frequently utilised for integration site identification include the linker mediated polymerase chain reaction (LM-PCR), the *Alu*-HIV assays and the Integration Site Loop Amplification (ISLA) assay. The LM-PCR approach ligates a DNA linker to the ends of fragmented DNA and selectively amplifies the junctions between the host and viral DNA using primers that are specifically designed for viral DNA (often positioned in LTR) and the linker. Recent improvements to this assay have been described by Wells and colleagues in 2020, where DNA sample fragmentation, end-repair and linker-ligation are completed in a single enzymatic random cleavage step which replaced the original multistep process that utilised

mechanical shearing by a double-stranded DNA fragmentase to fragment the DNA followed an independent linker-ligation step. Overall, this assay targets the LTR ends of the HIV genome and 5–15% of HIV integration sites are recovered with the improved method (Wells *et al*., 2020). The *Alu*-HIV assays are more widely utilised than LM-PCR methods, and the real-time *Alu*-PCR assays have been used to quantify integrated HIV relative to unintegrated HIV. The first step utilises a forward primer which binds to the human *Alu* element, a repeat element that is encountered every 3 000 base pairs, and a HIV specific reverse primer. Different HIV specific primers have been utilised by different research groups and are chosen based on the research question. Both an *Alu*-Gag (*Alu* and HIV Gag specific primers) and an *Alu*-LTR (*Alu* and HIV LTR specific primers) (Vandergeeten *et al*., 2014) approach have been described. Comparisons between the two approaches have shown that the *Alu*-LTR approach is capable of capturing all integrated HIV DNA thereby increasing the approach's sensitivity. The second step in the assay utilises a real-time PCR that targets HIV LTR. The distances between the *Alu* element and the integrated provirus varies for each infected cell and different amplification efficiencies are encountered which are distance dependent. In order to overcome the variations in efficiency, repeated sampling is required (Pinzone & O'Doherty, 2018). Significant weaknesses are also observed in this approach as integration sites linked to defective proviruses cannot be distinguished from intact and potentially replication-competent proviruses. The ISLA assay developed by Wagner and colleagues in 2014 is a unique approach that was designed to identify integration sites in single cells and determine whether integration into specific genes may promote the proliferation of HIV-infected cells. The assay utilised two forward primers in HIV *env* and HIV *nef* to linearly amplify the proviral DNA and the adjacent integration site. A random priming step followed where oligos with 3' random decamers and a 5' reverse complement U5-specific primer sequence were used to incorporate a self-complementary tail, which after complementary strand synthesis and denaturation, would loop back onto the complementary U5 sequence of the HIV template. The generated loop contained the human genomic sequence with complementary HIV ends. Three rounds of exponential amplification followed and the templates were sequenced to identify the integration sites. The proviral integrity was not confirmed with this assay as the ends of the HIV genome were targeted for integration site identification. The molecular-based integration site assays as described here and published by Madarelli (2014) and Cohn (2015) randomly fragment the sample to investigate HIV integration sites. And although these methods allow for the recovery and identification of integration sites, random fragmentation disrupts the HIV genome and prevents simultaneous identification of the proviral integrity and the respective integration

site. Further challenges encountered with integration site assays and analyses include that direct evidence for clonal expansion is provided but the replication competence of the provirus cannot be established.

The standard method for identifying replication-competent proviruses is investigated through a culture-based method known as a viral outgrowth assay (VOA). Although replication competence can be inferred through this method, identifying the clonality of these viruses poses further problems. Multiple rounds of *in vitro* infection are required in VOA cultures and the HIV integration sites that are captured are no longer a true reflection of those that occur *in vivo* (Bui *et al*., 2017, Lorenzi *et al*., 2016, Hosmane *et al*., 2017).

Additionally, full-length HIV proviral sequencing methods aimed at identifying clonally expanded HIV, amplify regions spanning the HIV genome but exclude information on integration sites and replication competence (Ho *et al*., 2013, Bruner *et al*., 2016). The integration sites of intact replication-competent or potentially replication-competent proviruses need to be efficiently and simultaneously obtained to bypass the problems encountered by the methods mentioned above, which creates a technical challenge.

Several methods have been developed to simultaneously identify intact HIV proviruses and integration sites. Similar methods were developed by Einkauf *et al*. (2019) and Patro *et al*. (2019) to identify expanded clones of cells from donors infected with HIV and utilise a combined approach to characterise proviral integrity and the respective integration sites. Both methods utilised a whole genome amplification approach using the isothermal phi29 polymerase enzyme to generate 1 000 to 10 000 copies of an individual provirus. The study by Einkauf and colleagues (2019) showed that three donors carried expanded clones and a portion of clones contained intact proviruses matching viral outgrowth in culture. Similarly, Patro and colleagues (2019) showed that the majority of the clones were defective in the five donors investigated except for one, which was identified as intact and was further confirmed to be replication-competent as the proviral sequence was also recovered by VOA. A limiting dilution culture and CD3/CD28-mediated proliferation method was utilised by the Siliciano group to sequence HIV integration sites and the proviral genome, however, only defective proviral clones were identified as viral cytopathic effects killed cells harbouring replication-competent provirus as a result of maximum T cell activation (Liu *et al*., 2020). Overall, the methods described by Einkauf (2019) and Patro (2019) allow for simultaneous identification of rare intact HIV proviruses and integration sites, however, both methods

utilise a multiple step process of confirmation and are laborious, time-consuming and expensive. An efficient and cost-effective method for identifying, enriching and linking intact HIV proviruses and their integration sites is required as more than 50% of cells harbouring infectious HIV proviruses are maintained by clonal expansion.

1.7. Sequencing generations and technologies

Various sequencing platforms and chemistries have been used to facilitate investigations of clonal expansion as a mechanism of latent HIV reservoir persistence. DNA was first sequenced by Sanger sequencing in 1977 (Sanger *et al*., 1977). However, low throughput and increased costs were encountered when long stretches of DNA for larger-scale projects were sequenced, which brought about the next generation sequencing era which aimed to address these issues. Second-generation sequencing also known as massively parallel sequencing was developed as a means to cheaply generate millions of sequences per reaction per run in a short period of time. Second-generation sequencing includes various sequencing instruments developed by Roche, Applied Biosystems, Ion Torrent and Illumina®. Each sequencing platform utilises unique sequencing chemistries and methods to generate libraries which are presented in Table 1.1. The generation of short read lengths in all of the second-generation sequencing platforms is a significant drawback. Overall. no second-generation sequencing approach provides reads longer than 1 000 base pairs as shown in Table 1.1 (Weirather *et al*., 2017). Although these sequencing platforms are cost-effective, accurate and are supported by a wide range of analysis tools and pipelines, the generation of short reads make resolving repetitive regions in genomes difficult, as misassembles and gaps are regularly encountered (Rhoads & Au, 2015, Amarasinghe *et al*., 2020). In addition, large structural variations are more challenging to detect and characterise with these sequencing platforms. Second-generation sequencing heavily relies on PCR to enrich the DNA template and regions with high GC% are inefficiently amplified (Lu *et al*., 2016).

Third-generation sequencing overcomes the limitations identified in second-generation sequencing as it offers the ability to sequence single nucleic acid molecules in real-time and generates long read lengths at high resolution. Two sequencing platforms currently dominate this market, namely Pacific Bioscience's (PacBio®) single-molecule, real-time (SMRT) sequencing and Oxford Nanopore Technologies' (ONT) nanopore sequencing which were commercially released in 2011 and 2014, respectively (Amarasinghe *et al*.,

2020). A major advantage of these technologies is that the native DNA or RNA is sequenced which eliminates biases that are potentially introduced by PCR amplification, thereby preserving base modifications. The generation of long reads further improves *de novo* assemblies of genomes that include large stretches of repetitive regions, mapping certainty, detection of structural variants, transcriptome analysis and also plays an important role in metagenomic research (Rhoads & Au, 2015, Lu *et al*., 2016, Amarasinghe *et al*., 2020).

Major technical differences separate PacBio®'s SMRT sequencing and ONT's nanopore sequencing, however, both sequencing technologies are capable of directly sequencing single nucleic acid molecules and generating long reads of between 1 – 100 000 base pairs in length (Lu *et al*., 2016, Weirather *et al*., 2017, Krishnakumar *et al*., 2018). Further differences between the two sequencing technologies are highlighted in Table 1.1 and the sequencing approaches are briefly explained in sections 1.7.1 and 1.7.2.

Table 1.1: Comparison of the sequencing technologies offered in the three sequencing generations

Sequencing generation	**Platform/ *devices***	**Library preparation**	**Chemistry/ sequencing method**	**Read accuracy**	**Maximum and average read length** (base pairs)	**Limitations**	**Strengths**
First	Sanger	PCR and cloning	Chain termination using ddNTPs	99.999%	≤1 000	• Not cost-effective for large stretches of DNA and larger scale projects • Low throughput and scalability	• Low cost instrument • Simple workflow • Gold standard method used for sequence confirmation
Second	Roche 454 *GS FLX+* *GS Junior*	Emulsion PCR	Pyrosequencing - cleavage of released pyrophosphate	99.997% (*GS FLX+*) >99% (*GS Junior*)	≤800 Average 400	• High reagent cost • High error rate in homopolymer repeats • Discontinued	• Longest read lengths of second-generation sequencers • Low error rate • Low nucleic acid input required
	Applied Biosystems SOLiD™	Emulsion PCR	Sequencing by ligation of hybridising labelled oligos	99.99%	≤100 Average 50	• Short read lengths • Long run time (7 -8 days)	• Low cost instrument • Low error rate
	Ion Torrent®	Emulsion PCR	Ion semiconductor sequencing	~98,5%	≤400 Average 200	• Short read lengths • High error rate in homopolymer repeats • Low scalability	• Low cost instrument • Short run time • Low nucleic acid input required

	Illumina® *HiSeq™* *MiSeq™*	Bridge PCR	Sequencing by synthesis using reversible terminators	>99%	2x 125 2x 250 Paired end reads	• Short read lengths • Low multiplexing capabilities • High instrument cost	• High throughput • Low error rate • Low nucleic acid input required • Lowest cost per base • Direct sequencing of samples for some applications
Third	Pacific Biosciences *RSII* *Sequel* *Sequel II*	No PCR required Hairpin adaptor ligation generates a circular sequencing template	Single-molecule real-time sequencing	>99%	≤30 000 Ultra-long reads	• High error rate (overcome by SMRTbell* resequencing) • High cost instrument • Read-length is limited by polymerase longevity	• Real-time measurement of base incorporation • Large number of reads per single molecule • Fast run times • Long read lengths • No amplification required
	Oxford Nanopore Technologies *MinION* *GridION* *PromethION*	No PCR required but can be used to increase low concentration DNA/RNA Sequencing adaptors are ligated to the ends of DNA or RNA molecules	Electrical conductivity	>97% Consensus accuracies: >99.99% for R9.4.1 >99.999% for R10.3	≤2 300 000 Ultra-long reads	• High error rate in variable regions and homopolymers • Very high concentration nucleic acids need to be prepared prior to sequencing	• Real-time sequencing with real-time data output • Large number of reads per single molecule • Long read lengths • No amplification required • Fast run times • Low instrument and reagent cost • Multiplexing capabilities • Portable and scalable

* SMRTbell is a double-stranded DNA template that has hairpin loops ligated at both ends, it forms the DNA library for Pacific Biosciences SMRT sequencing.

Sources: Ion Proton vs Illumina HiSeq2500 vs SOLiD 5500, 2013, Buermans & den Dunnen, 2014, Rhoads & Au, 2015, Illumina®, 2017, De Maio *et a*l., 2019, PacBio®, 2020, Oxford Nanopore Technologies, 2020.

1.7.1. PacBio® SMRT sequencing

SMRT sequencing detects fluorescence events that correspond to the addition of a fluorescently-labelled nucleotide by a polymerase that is tethered to the bottom of a small well surrounded by aluminium walls called Zero-mode waveguide (ZMW) measuring 70nm by 100nm with about 150 000 ZMW in one SMRT cell. The DNA polymerase in the ZMW binds to a DNA molecule and incorporates fluorescently-labelled nucleotides that emit light pulses. The replication process in all of the wells on the SMRT cell are recorded and the light pulses correspond to the sequence of the nucleic acid, which is also known as a continuous long read (CLR) (Pacific Bioscience, 2020, Quainoo *et al*., 2017). The process has a high read error, but this can be overcome by the use of the SMRTbell sequencing library approach: the sample is prepared by ligating hairpin adaptors to the ends of a double-stranded DNA fragment creating a circular template referred to as a SMRTbell. The circular template allows the polymerase enzyme to replicate both strands multiple times, increasing the accuracy of the sequencing read. CLRs can be split into multiple reads known as subreads by recognizing and removing the adaptor sequences. The consensus of multiple subreads in a single well yields a circular continuous sequence (CCS) read with higher accuracy as shown in Figure 1.6 (Rhoads & Au, 2015). The read length is dependent on polymerase longevity and library insert sizes range from 250 base pairs to 50 000 base pairs. However, larger inserts can prevent multiple replications of the circular template which affects the CCS quality. The average error rate of this third-generation sequencing technology is estimated at <1% (Amarasinghe *et al*., 2020). CCS reads retain errors and exhibit bias for insertions and deletions in homopolymer regions (Wenger *et al*., 2019). PacBio® constantly improves their sequencing technology by providing regular updates to hardware, chemistry and software in an attempt to compete with the sequencing accuracy achieved by second-generation sequencing technologies.

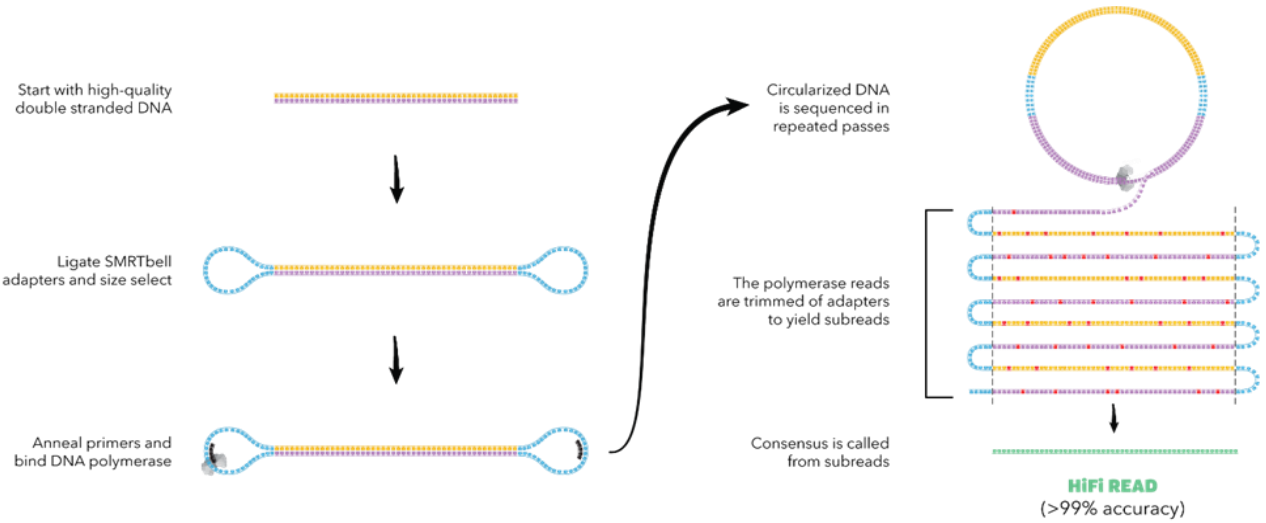

Figure 1.6: Library preparation, SMRT sequencing and the generation of CCS sequencing reads. (Source: PacBio®, 2020. [Online]: https://www.pacb.com/smrt-science/attachment/how-to-get-hifi-reads_v2/).

1.7.2. ONT nanopore sequencing

ONT's nanopore sequencing measures the changes in ionic current that occur across an electrically resistant polymer membrane as single-stranded nucleic acids pass through biological nanopores that are present on a flow cell. The nucleic acid sequence is inferred as each nucleotide confers a different resistance which is measured by an arrayed sensor chip and passed to an Application-Specific Integrated Circuit (ASIC) which controls and measures the experiments. The sequence data is processed by the MinKNOW specialist software that is responsible for carrying out several core functions which includes; data acquisition, real-time analysis and feedback, data streaming while providing device control, sample identification and tracking to ensure that the chemistry of the platform performs optimally while the samples are processed. ONT offers a customisable sequencing approach in which various sequencing devices with different throughputs and library preparation methods can be chosen based on the purpose of the experiment. Three sequencing devices are currently available, namely; the MinION (sequences one flow cell), GridION (up to five flow cells) and PromethION (up to 48 flow cells). In addition, three types of flow cells are available for use on these sequencing devices. The MinION and GridION devices use the FLO-MIN106D flow cell and the PromethION uses the FLO-PRO002 flow cell. The third flow cell is used with the Flongle which is an adaptor for the MinION or GridION and uses single-use FLO-FLG001 flow cells for small sequencing tests (Oxford Nanopore Technologies, Nanoporetech, 2020). Three different library preparation options can be utilised to ligate sequencing adaptors to the ends of DNA or RNA fragments. The two most commonly used library preparations known as the 1D and Rapid approaches, allow for the

template and complement strands to be individually sequenced whereas the 1D^2 library preparation uses special adaptors to increase the probability that the complement strand will immediately follow the template strand to increase read accuracy (Oxford Nanopore Technologies, 2018). The single pass accuracy of these library preparation protocols is 95% and 98%, respectively and an overall average accuracy of 97% (Amarasinghe *et al.,* 2020). ONT sequencing provides the longest read lengths of all sequencing technologies, with a current record length of 2 300 000 base pairs (Payne *et al*., 2019) and library inserts of 10 000 to 30 000 base pairs are commonly generated. Insertions, deletions and substitutions are frequently observed in ONT nanopore sequencing data which is influenced by the biological nanopore present on the flow cell. Resolving low complexity stretches and homopolymer sequences is difficult as the current that is measured is a function of the particular *k*-mer that resides in the nanopore at the time, and because translocation of homopolymers does not change the sequence of the nucleotides within the pore, it results in a constant signal that makes determining homopolymer length difficult. This problem was mostly encountered with the R9 nanopores however, the latest R10 nanopores increase accuracy over homopolymer regions (Brown, 2019). Similar to PacBio®, ONT regularly release new chemistry and software updates that improve read quality.

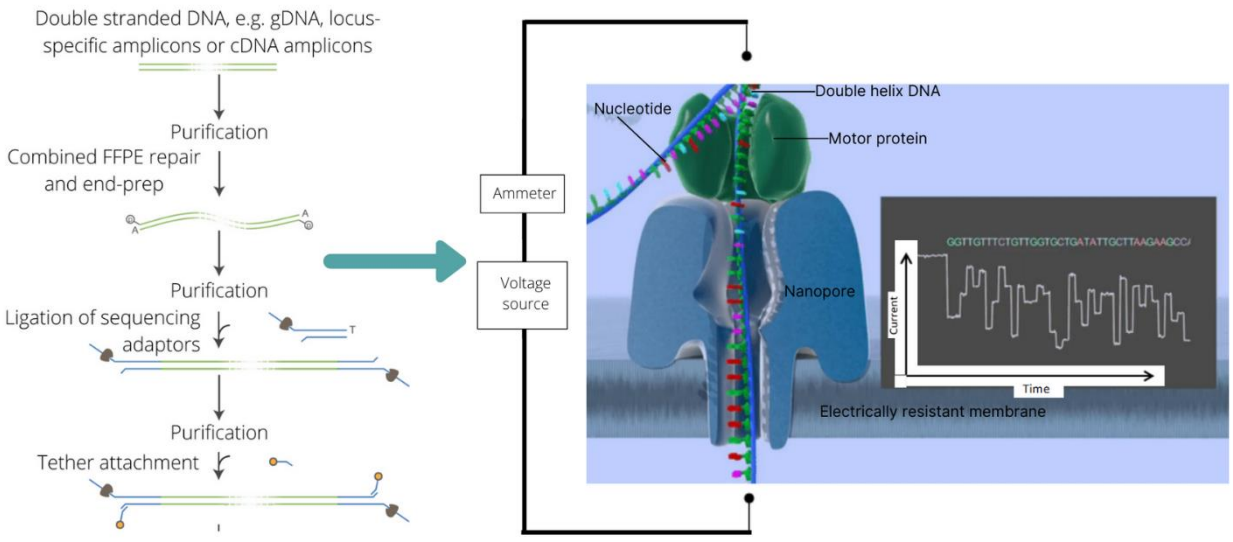

Figure 1.7: ONT library preparation and sequencing. (Adapted from sources: Oxford Nanopore Technologies: Amplicons by Ligation protocol, 2019 and Oxford Nanopore Technologies, 2017. [Online]: https://www.youtube.com/watch?v=GUb1TZvMWsw&feature=youtu.be).

ONT's nanopore sequencing offers several advantages over PacBio®'s SMRT sequencing, although the technology is still in active optimisation and has a higher reported sequencing error rate. All three sequencing devices offered by ONT are smaller than those developed

by PacBio® which offers portability. PacBio® platforms have relatively low operating costs when assessed on a per sample basis but the initial investment cost is much larger than for ONT devices. ONT offers a MinION starter kit at $1 000 which includes a MinION sequencing device, two flow cells and two preparation kits (Oxford Nanopore Technologies, Nanoporetech, 2020). Furthermore, ONT flow cells are reusable, provided sufficient active nanopores are available on the flow cell and multiplexing is possible which significantly lowers the cost per sample. ONT nanopore sequencing does not rely on sequence replication and therefore, is not limited by the longevity of an enzyme. The nanopores of the flow cell will proceed to read more DNA molecules until the life span of the nanopore is exhausted, current standard sequencing times are set to 72 hours. ONT has reported longer average read lengths than PacBio® which could facilitate the assemblies of longer genomes. Individual light signals are recorded in real-time with SMRT sequencing but data output can only be accessed if the run is stopped. In comparison, ONT's data output occurs in real-time which allows for almost real-time data analysis within 2 minutes of run initiation and the sequencing run can be terminated by the user when sufficient sequence coverage is obtained or when the desired mutation is detected in real-time (Quainoo *et al*., 2017).

A major gap in the field of HIV research presents itself, an efficient method to identify HIV integration sites of an intact proviral genome is currently not available. Third-generation sequencing may help overcome the current technical barriers and eliminate laborious and time-consuming methods that are currently utilised and may help facilitate the development of methods that are more efficient and allow for simultaneous investigation of these features of HIV. Long-lived HIV infected cells, which constitute HIV reservoirs are currently the most important barrier to achieving HIV cures. The field of HIV cure research aims to eliminate the latent reservoir, or prevent viral rebound from these reservoirs. The development and implementation of efficient methods to sequence full-length HIV proviruses and link these to integration sites, which uniquely define the cell clones that harbour these proviruses, would aid in the characterization of HIV reservoirs. This improved understanding of the nature of HIV reservoirs is pivotal for the advancement of the field of HIV cure research.

In this study, an assay was designed to efficiently link and sequence integration sites of provisionally intact proviral sequences, previously identified by Illumina MiSeq sequencing, belonging to HIV-1 subtype C as many protocols have been designed and optimised for HIV-1 subtype B. The third-generation sequencing platform developed by ONT was utilised to sequence these samples and near-full-length HIV-1 amplicons to determine the efficiency

of this sequencing technology in comparison to the reference Illumina® MiSeq™ sequencing platform.

1.8. Research question, Hypothesis, Aims and Objectives of the study

This is a descriptive retrospective design study.

1.8.1. Research questions

This study aimed to address the following questions:

1. Where do HIV proviruses preferentially integrate into the human genome?
2. What is the most effective way to sequence integrated and/or near-full-length (NFL) intact proviruses?

1.8.2. Hypotheses

The study is further driven by two hypotheses based on the research questions posed above:

1. HIV-1 integration sites can be recovered from pre-nested PCR products that yield intact or potentially intact proviruses.
2. ONT sequencing will provide an efficient approach to sequence NFL HIV-1 proviruses and their integration sites.

1.8.3. Aim of the study

To investigate the integration sites of proviral HIV-1 DNA in children with HIV early antiretroviral (CHER) and post-CHER cohort to determine preferential sites of integration in the human genome.

1.8.4. Specific study objectives

1. To establish and optimise an ISLA for HIV-1 subtype C.
2. To use the optimised assay to detect integration sites in pre-nested PCR products of intact or likely intact proviruses in the CHER/post-CHER cohort.
3. To sequence ISLA amplification products using Sanger sequencing to determine where the provirus integrates into the human genome.
4. To establish and determine if ONT sequencing is an effective sequencing platform for the detection of proviruses.

Chapter 2

2 Materials and methods

2.1. Ethical approval

This study was granted initial ethical approval on 20 September 2017 through the Health Research Ethics Committee of Stellenbosch University, ethics reference number M14/07/029. Prior to the lapse of the above-mentioned ethical approval, a new application was submitted on 15 February 2018 and granted approval on 27 March 2018, ethics reference number N18/02/020. Informed consent was obtained before patients were included in the study.

2.2. CHER and Post-CHER cohort background

Participants meeting the following criteria were recruited as part of a clinical trial and formed the CHER cohort; infants aged <12 weeks with a confirmed HIV infection by a positive PCR test, plasma HIV RNA viral load of >1000 copies/ml and a CD4 percentage ≥25%. Infants with CD4% >25% were randomised into three study arms with different treatment strategies. The three study arms were defined as follows; 1) deferred ART, where infants were not immediately initiated on therapy, 2) immediate ART treatment with a subsequent interruption at 40 weeks post-ART initiation and 3) immediate ART treatment with a subsequent interruption at 96 weeks post-ART initiation. The trial was conducted in two locations in South Africa which included the HIV Research Unit at Chris Hani Baragwanath Hospital in Johannesburg and the Children's Infectious Disease Clinical Research Unit at Tygerberg Hospital in Cape Town (Cotton *et al*., 2013, Violari *et al*., 2008). Following the conclusion of the initial trial, a subset of children were retained in a Post-CHER descriptive study to investigate neurocognitive outcomes and HIV-1 reservoirs.

2.2.1. Criteria of the participants used in this study

The inclusion criteria for participant selection will be briefly described and is detailed in Katusiime *et al*. 2020. Initially, eight participants enrolled in the Post-CHER follow up studies were selected based on having sufficient sample material available for investigation and to increase the likelihood of obtaining NFL amplicons, samples with HIV-cell associated DNA loads higher than 40 copies per million peripheral blood mononuclear cells (PBMCs) at the 6 to 7-year sampling period were used. Overall, nine provisionally intact NFL amplicons belonging to four participants were initially identified by Illumina MiSeq and investigated in

this study. Seven of these NFL amplicons were later confirmed to be intact while the remaining two were identified as non-intact as their sequences contained deletions. All of the participants maintained undetectable viral loads within the most recent 36 months of the sampling time point that was utilised for NFL amplification.

2.3. Methods

2.3.1. Project rationale

Intact HIV-1 NFL proviruses were previously identified in children belonging to the CHER and post-CHER cohorts (Katusiime *et al*., 2020), showing that these intact proviruses belong to $CD4^+$ T cell clones require a technique to link intact proviruses to their respective integration sites. The large majority of integration sites represent that of defective proviruses, therefore it is more efficient to identify intact proviruses first and then attempt to recover their integration sites. A method to achieve this is currently unavailable. The aim of this project was to develop a technique that efficiently recovers integration sites of known intact HIV-1 proviruses and to efficiently sequence NFL proviruses using the third-generation sequencing technology developed by Oxford Nanopore Technologies (ONT).

2.3.2. General molecular methods

2.3.2.1. Peripheral blood mononuclear cell isolation

PBMCs were isolated from ethylenediaminetetraacetic acid (EDTA) whole blood samples which were collected from patients belonging to the CHER and post-CHER cohorts. The isolation protocol used was adapted from the University of Pittsburgh (Brown & Buffo, 2007) and uses density gradient centrifugation and Histopaque®-1077 (Sigma-Aldrich, Missouri, USA) to remove erythrocytes and retain the PBMC population which include, B cells, T cells, Natural Killer cells, monocytes and dendritic cells. Isolated PBMCs were stored in cryopreservation media consisting of 10% dimethyl sulfoxide (DMSO) (Sigma-Aldrich, Missouri, USA) and 90% fetal bovine serum (FBS) (Thermo Scientific, Waltham, Massachusetts) determined by an average live count of the isolated PBMCs performed on Bio-Rad's TC20TM Automated Cell Counter (Bio-rad, California, USA). The PBMC's were stored in Mr Frosty™'s (Nalgene, New York, USA) at - 80°C for 24 hours and then moved into a rack in a liquid nitrogen (LN_2) tank.

2.3.2.2. Nucleic acid extraction

Viral nucleic acid was extracted from PBMCs using a method published by Hong *et al*. in 2016. PBMCs stored in cryopreservation media were removed from the LN_2 tank and allowed to equilibrate overnight at -80°C. Following the overnight step, the PBMCs were thawed in a bead bath at 37°C for 2 minutes. An equal volume of Roswell Park Memorial Institute (RPMI) Medium 1640 (BioWhittaker®, Lonza, Basel, Switzerland) heated to 37°C was added dropwise to thawed cryopreserved PBMCs. The PBMCs were pelleted by centrifugation at 500 RCF for 5 minutes and the supernatant was removed. One milliliter of Lysis buffer 1 consisting of 3 M guanidinium hydrochloride (GuHCl) (Sigma-Aldrich, Missouri, USA), 50 mM trisaminomethane hydrochloride (Tris-HCl) (Sigma-Aldrich, Missouri, USA) and 1 mM Calcium chloride ($CaCl_2$) (Sigma-Aldrich, Missouri, USA) was prepared. Prior to use, 50 µl of proteinase K (ProK) (20 mg/ml) (Ambion®, Texas, USA) was added to Lysis buffer 1 and 100 µl was added to the cell pellet and sonicated for 10 seconds using the Omni Sonic Ruptor-400 Ultrasonic Homogenizer (Omni International, Georgia, USA) in a Branson Sonifier ultrasonic cell disrupter cup horn (Branson, Connecticut, USA) at an amplitude of 60% in pulse mode. Following sonication, the sample was incubated for 1 hour at 42°C. A second buffer, Lysis buffer 2, consisting of 5.7 M guanidinium isothiocyanate (GuSCN) (Sigma-Aldrich, Missouri, USA), 50 mM Tris-HCl and 1 mM EDTA was prepared. Prior to use, 30 µl of glycogen (20 mg/ml) (Roche, Basel, Switzerland) was added to 1 ml of Lysis buffer 2 and 400 µl was added to the sample which was mixed and incubated for a further 10 minutes at 42°C. Five hundred microliters of room temperature 100% isopropanol (Sigma-Aldrich, Missouri, USA) was added to the sample, vortexed and centrifuged at 21 000 RCF for 10 minutes to pellet the nucleic acids. The supernatant was removed and the pellet was washed with 750 µl of 70% ethanol (Sigma-Aldrich, Missouri, USA), vortexed and centrifuged at 21 000 RCF for 10 minutes. After centrifugation, the supernatant was removed and the pellet was air-dried until nearly transparent in appearance. Lastly, the pellet was gently resuspended by centrifugation in 70 µl of 5 mM (Tris-HCl). Isolated viral nucleic acids were stored at - 80°C.

2.3.2.3. Single genome amplification

Single genome amplification (SGA) is based on limiting dilution and a highly sensitive nested PCR amplification method. In short, DNA is diluted to an endpoint at which 30% of replicates are positive. At this concentration, according to the Poisson distribution, each positive PCR reaction most likely represents the amplification of a single genome. The PCR methods

mentioned in sections 2.3.2.3.1. and 2.3.2.3.2. are SGA approaches to ensure that downstream methods and results are only as a result of a single genome. Extracted viral DNA is therefore serially diluted with cold 5 mM Tris-HCl, each dilution was done in replicate and viewed as mentioned in section 2.3.2.3.3.1. to identify which dilution is the target that gives a 30% positive amplification result. Once identified, viral DNA is diluted at the target and amplification occurs in replicates of 84 in a 96 well PCR plate. Four of the remaining wells on the plate were used to include controls, 2 wells were used as non-template controls and 2 wells were used as positive controls. The non-template controls contained no DNA template and functioned to ensure that no contamination occurred during the reaction set-up. The positive controls contained a DNA template that was known to contain the target and functioned to ensure that the primer set, reaction and thermocycling conditions were set up correctly.

2.3.2.3.1. Near-full-length HIV-1 proviral amplification

Extracted samples from patients meeting specific criteria mentioned in section 2.2.1 were subject to amplification using HIV near-full-length primer sets for the pre-nested and nested PCR reactions. Initial PCRs were conducted using published primer sets from Li *et al*. (2010). Pre-nested reactions used primer set Li_OuterF (5'-AAA TCT CTA GCA GTG GCG CCC GAA CAG-3') and Li_OuterR (5'-TGA GGG ATC TCT AGT TAC CAG AGT C-3') and nested reactions used primer set Li_InnerF (5'-GCG GAG GCT AGA AGG AGA GAG ATG G-3') and Li_InnerR (5'-GCA CTC AAG GCA AGC TTT ATT GAG GCT TA-3'). Amplification using these primer sets would produce an amplicon size of 8 700 base pairs (8.7kb). The nested primer set was later changed to include Li_OuterR and a newly designed primer published in Katusiime *et al*. (2020) NFL_alt_in_F (5'-CCG AAC AGG GAC BHG AAA GCG AA-3') to amplify an important region of HIV that included the packaging signal, as this was recently shown to be essential for replication competence (Das *et al.,* 2019). Positive amplification using these primers resulted in an amplicon size of 9 000 base pairs (9kb).

The pre-nested PCR reaction was performed in a 0.2 ml 96 well PCR plate, to which 5 µl of 2X RANGER Mix from Bioline's RANGER Mix kit (Bioline, London, United Kingdom) were added. In addition to the aforementioned, 0.4 µl of both the forward and reverse primer at a concentration of 10 µM each and 2.2 µl of nuclease-free water were added. Two microliters of serially diluted or target diluted DNA template was added to the reaction mixture bringing the total reaction volume to 10 µl. Pre-nested amplicons are diluted with 83 µl of cold

Tris-HCl prior to the nested PCR reaction. Two microliters of the diluted pre-nested amplicons were added to their respective wells containing the nested PCR reagents. These reagents and their respective concentrations were used in the original and adapted NFL PCR approaches for both the pre-nested and nested reactions.

The PCR programme that was utilised for the pre-nested reaction consisted of an initial denaturation step at 95ºC for 2 minutes followed by 30 cycles, each consisting of cycling through 98ºC for 10 seconds and 61ºC for 10 minutes, followed by a final elongation step at 72ºC for 10 minutes. The original nested PCR programme used an initial denaturation step at 95°C for 2 minutes followed by 30 cycles, cycling through denaturation at 98°C for 10 seconds and an annealing and extension step at 65.5°C for 10 minutes, this was followed by a final extension at 72°C for 10 minutes. The adapted NFL nested PCR programme used an annealing temperature of 61.5°C instead of 65.5°C, but the duration and temperature of the remaining steps remained the same.

2.3.2.3.2. p6-PR-RT HIV-1 amplification

The presence of viral HIV DNA in extracted samples was determined using a published qualitative nested PCR method that amplified the p6 in *gag*, protease and reverse transcriptase (p6-PR-RT) regions of the HIV genome as shown in Figure 2.1. The pre-nested reaction used primer set, 1849+ (5'-GAT GAC AGC ATG TCA GGG AG-3') and 3500- (5'- CTA TYA AGT CTT TTG ATG GGT CAT AA-3') and the nested reaction used primer set, 1870+ (5'-GAG TGT TGG CTG AGG CAA TGA G-3') and 3410- (5'-CAG TTA GTC GTA CTA TGT CTG TTA GTG CTT-3') (van Zyl *et al.*, 2017). Successful amplification with the above-mentioned primers yield a product of 1 200 base pairs (1.2kb) in size.

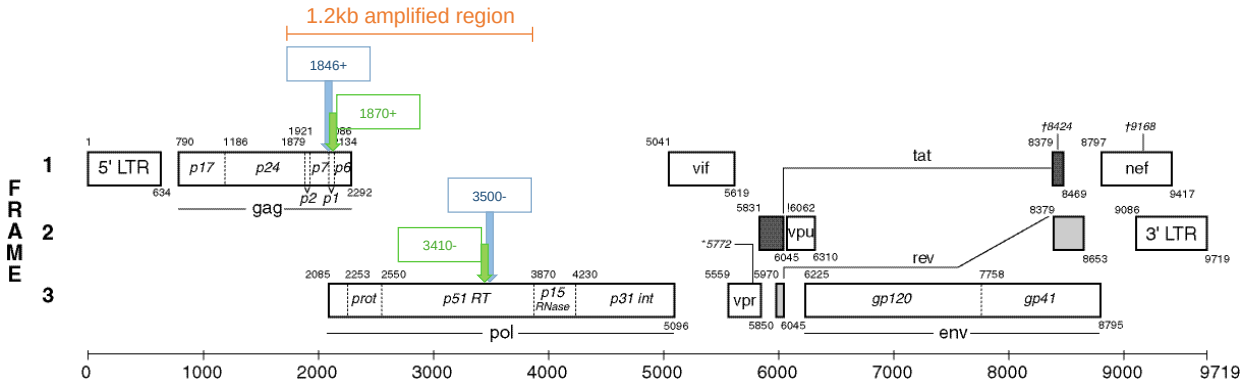

Figure 2.1: HIV-1 HXB2 reference genome indicating the primer binding sites and area amplified by the p6-PR-RT PCR primer sets. (Source: Los Alamos National Laboratory, 2020).

The PCR reaction was performed in a 0.2 ml 96 well PCR plate, to which reagents from Invitrogen's™ Platinum™ Taq DNA Polymerase High Fidelity kit (Invitrogen, Carlsbad, California, USA) were added at the supplied concentrations. This included 1 µl of 10X High Fidelity PCR Buffer, 0.4 µl of 50 mM $MgSO_4$ and 0.08 µl of 5 U/µl Platinum® Taq DNA Polymerase High Fidelity enzyme. In addition to the aforementioned, 0.2 µl of Invitrogen's dNTP Mix supplied at a concentration of 10 mM of each dNTP was added. Each primer was added at a volume of 0.04 µl and a concentration of 50 µM respectively. Lastly, 6.24 µl of nuclease-free water and 2 µl of serially or target diluted DNA template was added to each well for the pre-nested PCR reactions. For nested PCR reactions, 2 µl of pre-nested amplicons diluted with 83 µl cold 5 mM Tris-HCl were transferred to a plate containing the nested primer set, maintaining their alphanumerical position on each plate. For both pre- nested and nested PCR reactions the reagent volumes and concentrations remained the same.

The PCR programme that was utilised for the pre-nested PCR reaction consisted of an initial denaturation at 94°C for 2 minutes followed by 35 cycles, each consisting of cycling through 94°C for 30 seconds, 50°C for 30 seconds (annealing step) and 72°C for 3 minutes (extension step) followed by a final extension step at 72°C for 3 minutes. The nested PCR programme consisted of an initial denaturation at 94°C for 2 minutes followed by 40 cycles of 94°C for 30 seconds, 55°C for 30 seconds and 72°C for 1 minute. A final extension of 3 minutes at 72°C completed the PCR programme.

2.3.2.3.3. Viewing of single genome amplified products

To determine whether successful amplification has occurred, two different viewing approaches were utilised. The first approach is called a plate view (2.3.2.3.3.1.) and the second approach is the standard gel electrophoresis method (2.3.2.3.3.2.).

2.3.2.3.3.1. Plate view of amplicons

Once the nested PCRs were completed the amplified products were diluted with cold 5 mM Tris-HCl, the NFL nested PCR was diluted 1:4 and the p6-PR-RT nested PCR was diluted 1:5. Five microliters of the diluted nested PCR product was transferred to a new 96 well plate.

EZ Vision™ DNA Dye (VWR Life Science, Pennsylvania, USA) was diluted 1:10 000 with 5 mM Tris-HCl in a reagent reservoir and 15 µl of this was added to each well of the plate containing 5 µl of diluted nested PCR product. The presence of amplified products that complexed with the dye were visualised under ultraviolet light and photographed using the UVIprochemi II D-77LS-26M instrument (Uvitec, Cambridge, United Kingdom). The wells that contain amplified products that have complexed with the EZ Vision™ DNA Dye fluoresce, however this method does not confirm that the product represents the DNA sequence of interest.

2.3.2.3.3.2. Gel electrophoresis of amplicons

Wells that fluoresced with the plate view method described in the section above (2.3.2.3.3.1.) were subject to separation by gel electrophoresis to determine if the amplicon of interest was present. Ten microliters of amplified product was mixed with 1 µl of EZ Vision™ DNA Dye and was loaded into the well of a 1% agarose gel. Five microliters of Promega's 1kb DNA Ladder (Promega, California, USA) was loaded into the first and last well of each gel to serve as a reference for determining amplicon size. Gel electrophoresis was performed with the ENDURO™ Gel XL Electrophoresis System (Labnet, New Jersey, USA) by applying a voltage of 90V for 1 hour and 30 minutes using 1X SB buffer. The results of gel electrophoresis were visualised by exposure to ultraviolet light and photographed using the UVIprochemi II D-77LS-26M instrument.

2.3.3. Integration site loop amplification assay

2.3.3.1. Rationale

A technique to efficiently detect the integration sites linked to intact proviruses is not currently available. Intact proviruses are very rare, therefore, first identifying them and then attempting to obtain the linked integration site from linearly amplified products produced during a first round (pre-nested) PCR reaction (sections 2.3.2.3.1. and 2.3.2.3.2.) would provide the most efficient method of linking integration sites and intact proviruses.

2.3.3.2. HIV-1 subtype C integration site loop amplification assay

The ISLA protocol published by Wagner *et al*. in 2014 provided a method that could amplify across integration sites from linear amplification products and allow for the detection of HIV integration sites. The published method was optimised for the detection of HIV-1 subtype B integration sites however, as HIV-1 subtype C is the most prevalent subtype in Africa, the protocol was modified to target this subtype.

The ISLA assay was initially attempted using the pre-nested products from a NFL amplified patient sample that potentially contained an intact proviral HIV-1 sequence, identified by a PhD student in the research group at the time. However, due to the rarity of intact proviral HIV sequences and the limited number of patient PBMCs remaining for specific time points, the DNA amplified to generate the starting template for ISLA was changed. To further develop and optimise the ISLA assay, nucleic acid was extracted as mentioned in section 2.3.2.2. from the human lymphoblastic leukaemia cell clone 8E5/LAV (8E5) which contains a single integrated HIV-1 provirus and results in the production of defective virions (Folks *et al*., 1986, Quillent *et al*., 1993). The cells were originally stored at a concentration of 1.25 million cells per millilitre. Prior to use the concentration of the extracted DNA was measured using a DS-11 FX Series Spectrophotometer/Fluorometer (Denovix, Delaware, USA) and was initially diluted 1:100 with 5 mM Tris-HCl to start with 125 copies of HIV-1. A 2- or 3-fold serial dilution of the initial 125 copies of HIV-1 DNA was performed for the screening plates where the 8E5 template was diluted in HIV negative DNA or 5 mM Tris-HCl. At each dilution, each of 12 replicates were tested with a nested PCR reaction to determine the presence of an HIV genome. The final target dilution was chosen to achieve a 30% positivity rate, which according to the Poisson's distribution, would most likely represent a single genome in each replicate. A negative control and a non-template control were included on the plates where the provisionally intact HIV-1 patient sample and the 8E5

diluted in HIV negative DNA were amplified. The negative control contained DNA from HIV uninfected PBMCs as a template whereas the non-template control contained nuclease-free water. Further ISLA optimisations were conducted on 8E5 diluted in 5 mM Tris-HCl to target.

Step 1: *Linearly amplified products from a pre-nested PCR reaction served as a template for ISLA subtype C*

A PCR is typically used to exponentially amplify a specific segment of DNA or a gene however, PCR is also capable of generating linear-amplified templates in the background as shown in Figure 2.2.

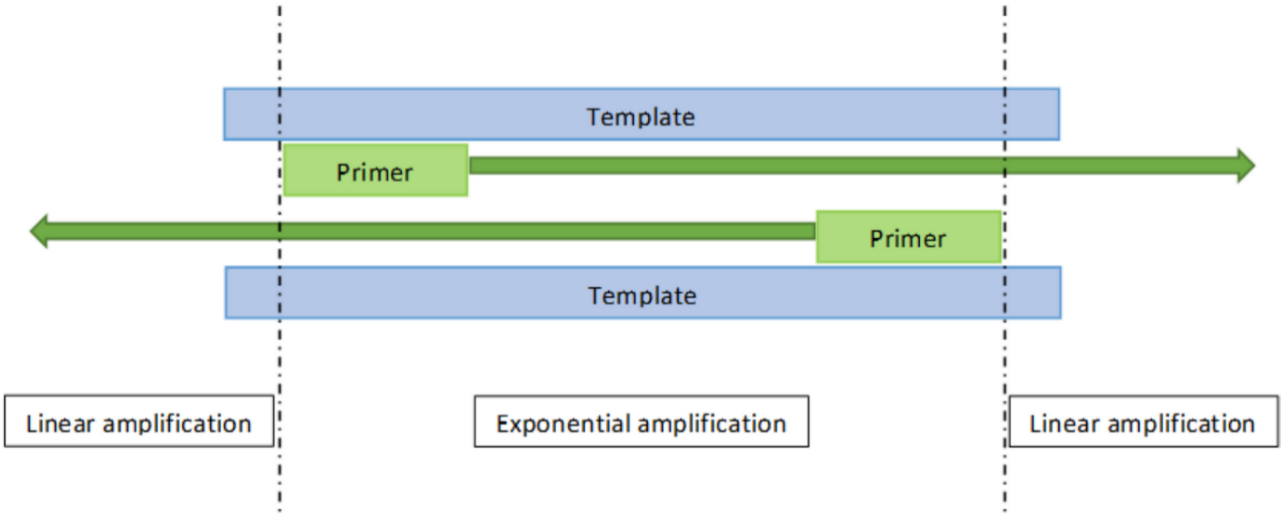

Figure 2.2: Overview of the production of linear amplified products during PCR.

Products from NFL PCR (section 2.3.2.3.1.) pre-nested reactions that underwent linear amplification served as a template for the initial attempts of ISLA as depicted in Figure 2.3.

Step 2: *Random priming with nanomers tailed with a U5-specific primer sequence*

A reverse primer site was created in this step by hybridising random nanomers tailed with a U5-specific sequence with an aliquot of the linear amplified single-stranded template. This step played an integral part in allowing for the formation of a loop containing the site of integration and the adjacent human genomic sequence which is described in *Step 4* of ISLA. The reaction consisted of adding 5 µl of 2X RANGER Mix, 4 µl of 10 µM random nanomer and 2 µl of nuclease-free water to a 0.2 ml PCR tube. Lastly, 9 µl of undiluted pre-nested product was added to the reaction.

The reaction was incubated at 68°C for 2 minutes, cooled to 65°C for 1 minute followed by gradual cooling 1°C per minute until reaching 25°C. The reactions were then reheated to 60°C and gradually cooled by 1°C per minute until reaching 20°C. The reaction was stored at -20°C if *Step 3* was not continued immediately.

Step 3: *Single-stranded fragments were digested by Exo I and filled with Taq DNA polymerase*

Exonuclease I (*Exo I*) (New England Biolabs, Massachusetts, USA) digested single-stranded DNA present on the 3' end of the random primer binding site (indicated with <---- in Figure 2.3). Ten units of *Exo I* was directly added to the reaction from *Step 2*.

The reaction was incubated at 37°C for 45 minutes, the temperature was then increased by 1°C every 3 minutes until reaching 43°C to ensure complete digestion by *Exo I* and to denature template that partially annealed to the U5 region of the primer. An incubation at 72°C for 15 minutes followed to allow *Taq* polymerase present in RANGER Mix to extend the region complementary to the U5-specific primer sequence indicated with ----> in Figure 2.3 and this was followed by a final incubation at 95°C for 5 minutes to inactivate *Exo I*.

Step 4: *Formation and extension of the loop containing the integration site and the adjacent human genome sequence*

This step is divided into 3 sub-steps to simplify the process of loop formation and comprises *Steps 4a*, *4b* and *4c* depicted in Figure 2.3, respectively. *Step 4a*, the previous steps allowed for the 3' end to be processed in such a way that the generated reverse complement sequence (indicated with ----> in *Step 4a* in Figure 2.3) could bind to its complementary site in U5 and form the loop. *Step 4b*, the addition of *Taq* polymerase present in RANGER Mix would result in the extension from the U5 priming site through the end of the fragment thereby creating a loop that contains the human genomic sequence with a HIV LTR that is duplicated on each end once the loop has formed. Primer RF2 binds and initiates exponential amplification as shown in *Step 4c*. This reaction was carried out in a 0.2 ml PCR tube to which the following reagents were added; 3 µl of 10 µM primer RF2, 25 µl of 2X RANGER Mix, 2 µl of nuclease-free water and 20 µl of the product formed in Step 3.

The thermocycling parameters of this step of ISLA included an initial denaturation at 95°C for 2 minutes followed by an initial 10 cycles cycling through, 94°C for 20 seconds, 60°C for 30 seconds and 72°C for 2 minutes; followed by 40 cycles cycling through, 92°C for 10 seconds, 65°C for 15 seconds and 72°C for 2 minutes and a final extension at 72°C for 5 minutes.

Step 5: *Nesting of the first round PCR products*

The first round PCR products were exponentially amplified by a second round of PCR amplification. Three microliters of 10 µM primer RF1, 25 µl of 2X RANGER Mix and 17 µl of nuclease-free water were added to a 0.2 ml PCR tube. Lastly, 5 µl of product from *Step 4* was added to the reaction.

The thermocycling parameters for this step consisted of an initial denaturation at 95°C for 2 minutes followed by 34 cycles cycling through, 94°C for 20 seconds, 65°C for 30 seconds and 72°C for 2 minutes with a final extension at 72°C for 5 minutes.

Step 6: *Visualisation and Sanger sequencing of ISLA amplified products*

ISLA amplified products measuring approximately ≥300 base pairs in length were identified through separation by gel electrophoresis. Sufficient stretches of HIV and the human genomic sequence contained in the loop were necessary to allow for the identification and linkage of the intact proviral sequence and the respective integration site. Products with an estimated base pair length of less than 300 would not allow for this identification as approximately 180 base pairs of the amplified product is designated to the duplicated LTR regions on both ends (approximately 90 base pairs each) and the remaining base pairs do not allow for a sufficient stretch of human genome to accurately determine the site of integration. ISLA products meeting the criteria of ≥300 base pairs were purified and sequenced by Sanger sequencing.

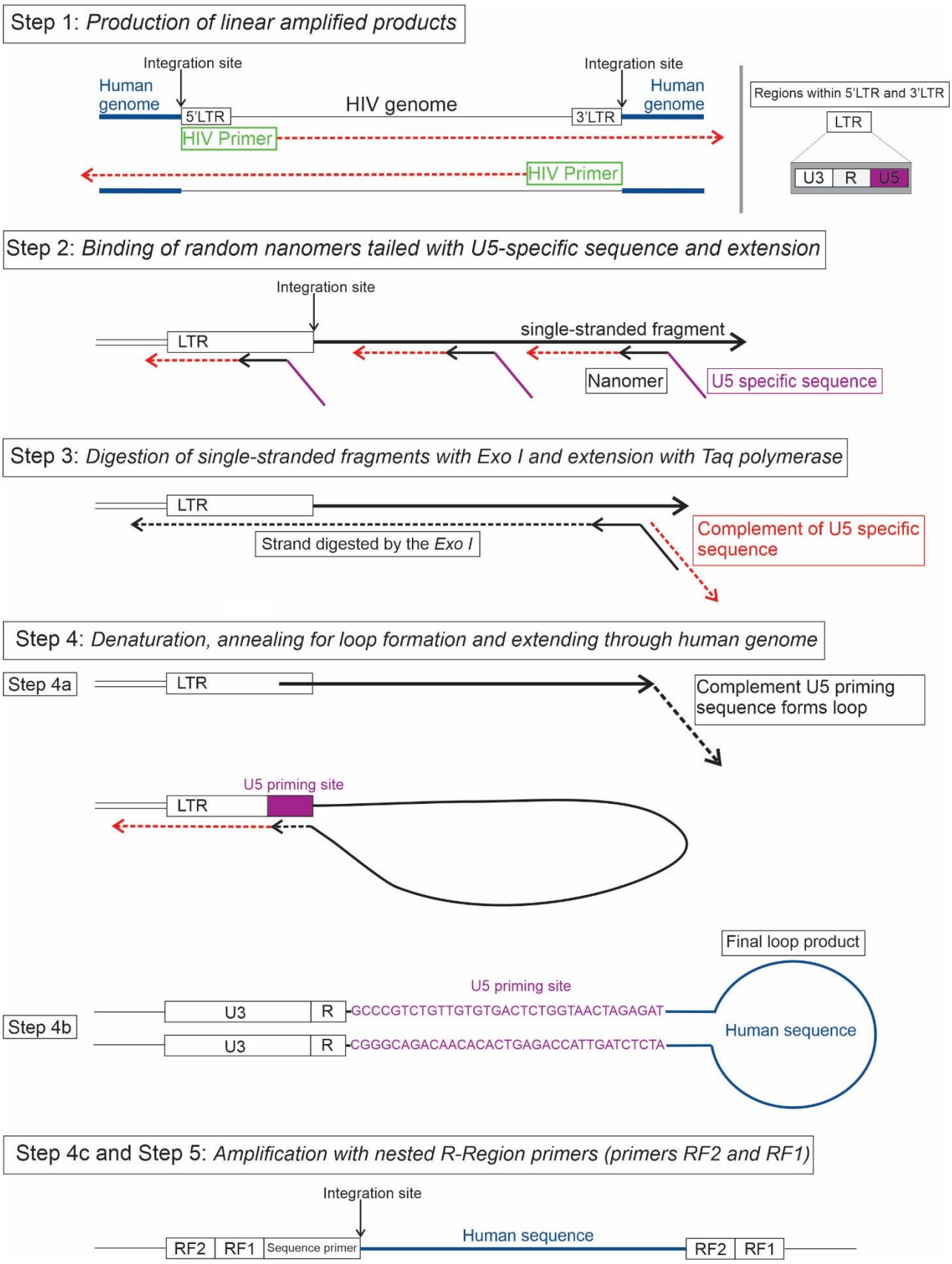

Figure 2.3: Subtype C ISLA approach outline adapted from the Wagner *et al*. (2014) supplementary materials. ----> indicate strands being generated in the current step. ----> indicate strands that have been generated in the previous step.

The protocol was further adapted to include independent amplification from either the 3'LTR or the 5'LTR end of the HIV genome and is explained in sections 2.3.3.2.2 and 2.3.3.2.3. HIV-1 subtype C specific primers were designed for these separate approaches.

2.3.3.2.1. Generating linear-amplified products for ISLA

Initial attempts at ISLA used linearly amplified NFL PCR products from pre-nested reactions (section 2.3.2.3.1.). Both the 3'LTR and 5'LTR ISLA approaches used these products as a template.

The NFL PCR primers overlapped with the primers used in ISLA and the approach to generating the starting template was adapted and changed, as shown in Figure 2.4, to the p6-PR-RT approach as no overlapping was present between the p6-PR-RT and ISLA primers.

2.3.3.2.2. ISLA assay from the 3' LTR end of HIV-1

The 3'LTR approach was attempted first as the NFL PCR enriched for the 3' end of the HIV-1 genome.

Table 2.1: Primers used in the 3'LTR ISLA approach and the respective steps

Primers 3'LTR amplification		
ISLA step	**Primer Name**	**Primer Sequence (5'-3')**
2	3'LTR-RP	GTG CCC GTC TGT TGT GTG ACT CTG GTN NNN NNN NN
4	3'LTR-RF2	AGA TCT GAG CCT GGG AGC TCT C
5	3'LTR-RF1	TTA AGC CTC AAT AAA GCT TGC CTT

2.3.3.2.3. ISLA assay from the 5' LTR end of HIV-1

This approach was attempted after unsuccessful amplification resulted from the 3' LTR ISLA approach.

Table 2.2: Primers used in the 5'LTR ISLA approach and the respective steps

Primers 5' LTR amplification		
ISLA step	**Primer Name**	**Primer Sequence (5'-3')**
2	5'LTR-RP	CCT GTG TGT GAT AGA CCC ACA AAT CNN NNN NNN N
4	5'LTR-RF2	CAA GGC AAG CTT TAT TGA GGC TTA A
5	5'LTR-RF1	GAG AGC TCC CAG GCT CAG ATC T

2.3.3.2.4. Determining optimal annealing temperatures for *Step 4* and *Step 5* of ISLA

The majority of the unique products produced by the initial ISLA attempts were <300 base pairs in length. The annealing temperatures of the primers utilised in *Steps 4* and *5* of ISLA were investigated as optimal annealing temperatures are required to avoid primer misbinding and also aid in yielding products of the required size to successfully identify integration sites. Therefore, two annealing temperatures were simultaneously investigated to ensure that the products being amplified were treated under the same conditions prior to PCR amplification. Lower annealing temperatures were selected based on the melting temperatures (T_m) of the primers.

The initial annealing temperatures of 60°C (10 cycles) and 65°C (40 cycles) were compared to the lowered annealing temperatures of 57°C (10 cycles) and 61°C (40 cycles) for *Step 4* of ISLA. *Step 5* of ISLA included comparing the initial annealing temperature of 65°C to a lowered annealing temperature of 53°C.

2.3.3.2.5. Increasing the input concentration of linear-amplified products for ISLA

A further adaption to the Wagner *et al*. (2014) method was the inclusion of a multiple displacement amplification (MDA) (shown in Figure 2.4) approach prior to *Step 2* of ISLA, based on personal communication with Dr Lisa Frenkel, who co-developed ISLA. This approach was added as the concentration of the input template was thought to be low as first round PCR products were used and the target, linear-amplified product was likely present at very low concentrations in comparison to exponentially amplified products. The MDA approach, which uses an isothermal enzyme, was used to amplify single-stranded fragments thereby increasing the concentration of the linear-amplified products.

Qiagen's REPLI-g® Mini Kit (Qiagen, Hilden, Germany) was used for the MDA step. The following reagents from Qiagen's REPLI-g® Mini Kit were added to a 0.2 ml PCR tube at the supplied concentrations; 14.5 µl of Reaction Buffer and 0.5 µl of REPLI-g Mini DNA Polymerase. For one MDA reaction, 5.5 µl of nuclease-free water and 4.5 µl of undiluted pre-nested PCR product was added to the reaction.

The MDA reaction mixture was incubated at 30°C for 16 hours with the thermocycler's lid temperature at 70°C followed by an enzyme inactivation step at 65°C for 3 minutes. The concentrations of these MDA amplified products were measured using NanoDrop™ 1000 Spectrophotometer (Thermo Scientific, Waltham, Massachusetts) to prevent adding a very high concentration of template to *Step 2* of ISLA. The concentration of the input templates needed to be at a concentration of ≤260 ng/µl. MDA produces hyperbranched structures which needed to be denatured prior to use in *Step 2* of ISLA. The branched products were incubated at 98°C for 5 minutes and snap cooled for 1 minute on ice to prevent single-stranded products from rebinding and forming branched structures again.

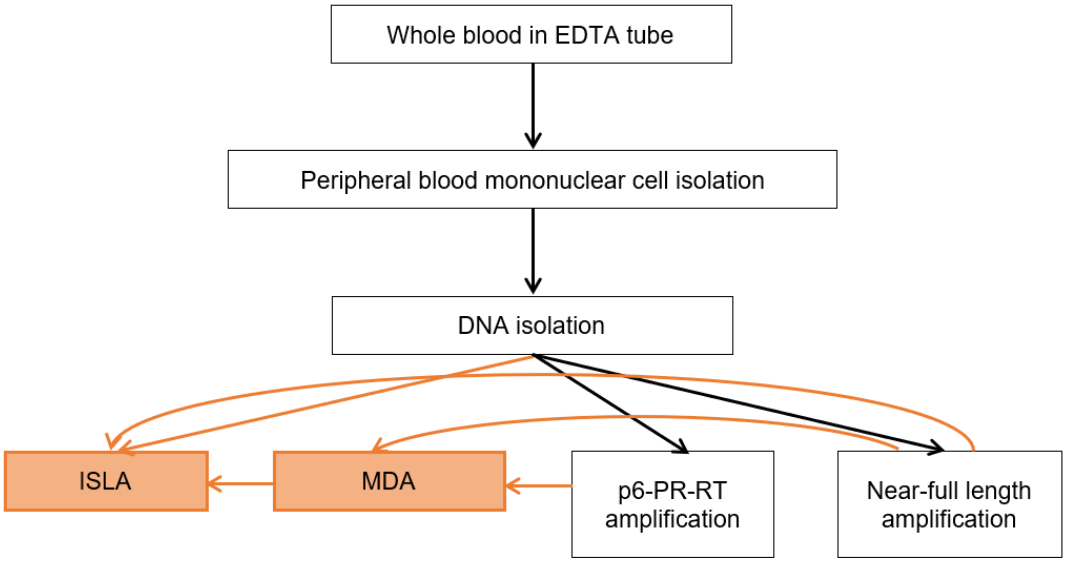

Figure 2.4: Diagram showing the change in the starting template for ISLA subtype C.

2.3.3.2.6. Investigating optimal random primer concentration

In an attempt to optimise *Step 2* of ISLA, a range of primer concentrations were investigated. The following concentrations of the random nanomer primer were investigated; 2 µM, 1.5 µM, 1 µM and 0.75 µM. The incubation temperatures and durations for *Step 2* remained the same for all primer concentrations investigated.

2.3.3.2.7. Comparison of different DNA polymerase enzymes

The protocol published by Wagner and colleagues used the MyTaq™ DNA polymerase enzyme (Bioline, London, UK) for the ISLA subtype B assay. The efficacy of RANGER Mix was compared to MyTaq™ Mix (Bioline, London, UK) which contained the MyTaq™ DNA polymerase enzyme in a ready-to-use mixture. MyTaq™ Mix replaced RANGER Mix in all ISLA steps and the incubation and PCR amplification steps remained the same as described in section 2.3.3.2.

2.3.3.2.8. Investigating the published subtype B ISLA assay

The published subtype B ISLA assay was investigated in our laboratory setting to determine its success in comparison to the modified subtype C ISLA assay. The subtype B ISLA assay was performed on 8E5 using the primers and methods described in the article published by Wagner *et al.* (2014).

Step 1 of the subtype B ISLA assay utilised a linear amplification step using two forward primers namely, ENV-WF (5'-CCC CGG CTG GTT TTG CGA TTC TAA AGT GTA-3') and NEF-WF (5'-CCA ATG CTG ATT GTG CCT GGC TAG AAG CA-3').

The reaction consisted of adding 1.5 µl of each forward primer at 10 µM concentrations, 5 U of *NotI*, 25 µl of RANGER Mix and 19.5 µl of nuclease-free water to a 0.2 ml PCR tube. Lastly, 2 µl of 8E5 template diluted to target (1:27) with Tris-HCl was added bringing the final reaction volume to 50 µl.

The cycling parameters used during Step 1 of the subtype B ISLA were as follows, 37ºC for 30 minutes, 95ºC for 2 minutes, followed by 30 cycles cycling through 94°C for 20 seconds, 60ºC for 30 seconds and 72ºC for 3 minutes.

Steps 2 to 5 used the primers described in Table 2.1. Furthermore, the incubation and thermocycling conditions for each step described in section 2.3.3.2. were used. An additional primer was included in this approach as described in the Wagner *et al.* (2014) method which served as a third round PCR amplification. Primer WF-1.U5 (5'-TCA AGT AGT GTG TGC CCG TCT GT-3') was used in this step and the thermocycling parameters were the same as described for *Step 5* of ISLA (section 2.3.3.2.).

2.3.3.3. Visualisation of ISLA assay products

All the wells that were amplified by the ISLA assay were initially visualised for the presence of positive amplification by using the plate viewing method as described in section 2.3.2.3.3.1. The sizes of the amplified products were visualised by separation through gel electrophoresis. The gel electrophoresis conditions were the same as described in section 2.3.2.3.3.2.

2.3.3.4. Purification of amplified products

Products successfully amplified with ISLA were purified for sequencing and cloning. Due to the presence of non-specific amplified products, products meeting the size requirement of ≥300 base pairs were identified by gel electrophoresis and wells containing products of the correct size were separated for a second time through gel electrophoresis on a 0.8% agarose gel, excised with a scalpel, placed into a 1.5 ml Eppendorf tube and subject to purification using Macherey-Nagel's NucleoTraP®CR kit for gel extraction following the manufacturer's instructions. The concentration and purity of the purified products were measured with the NanoDrop™ 1000 Spectrophotometer.

2.3.3.5. Sanger sequencing of ISLA assay products

Purified products were prepared for sequencing with reagents from the BigDye™ Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems™, California, USA). Prior to the sequencing reaction, the concentrations of the purified products were measured with the NanoDrop™ 1000 Spectrophotometer and diluted to have a DNA concentration measuring between 15 - 25 ng/µl. Reagents from Applied Biosystems'™ BigDye™ Terminator v1.1 Cycle Sequencing Kit were added to a 0.2 ml PCR tube; 1 µl of BigDye® Terminator v1.1 Ready Reaction Mix and 3 µl of 5X Sequencing Buffer at the supplied concentrations. The sequencing reaction for products generated by ISLA subtype C contained 1 µl of 5 µM 3'LTR Junction_seq (5'- TCTGGTAACTAGAGATCCCTCA -3') for the 3'LTR ISLA approach or Junction_seq (5'- CTT TTC TTR GAG TAA ATT AAC C -3') for the 5'LTR ISLA approach and the products generated by ISLA subtype B contained 1 µl of 5 µM WF-2.U5 (5'- GTT GTG TGA CTC TGG TAA CTA GAG AT-3'). Four microlitres of nuclease-free water and lastly 1 µl of purified product diluted to 20 ng/µl in nuclease-free water were added to the tube to bring the final volume of the reaction to 10 µl.

The thermocycling parameters that were used were as follows; 96°C for 0.10 seconds, 50°C for 0.05 seconds and 60°C for 4 minutes cycling through 25 cycles.

Plates containing these reactions were sent to the Central Analytical Facilities (CAF) at Stellenbosch University for sequencing clean-up and DNA sequencing.

2.3.3.6. Ligation of ISLA products in kit plasmid vectors

The purified ISLA products were ligated using the InsTAclone™ PCR Cloning Kit (Thermo Scientific, Massachusetts, USA) or the CloneJET™ PCR Cloning Kit (Thermo Scientific™, Massachusetts, USA) into the respective kit plasmid vectors. The InsTAclone™ PCR Cloning Kit was initially used for the ligation of ISLA products, the kit was discontinued while the project was ongoing and the CloneJET™ PCR Cloning Kit was used as the replacement.

2.3.3.6.1. Preparations of media and reagents prior to cloning

2.3.3.6.1.1. Luria broth

Sigma-Aldrich's Miller Luria broth (LB) powder medium (Sigma-Aldrich, Missouri, USA) was used to prepare LB culture medium for the growth of the transformed *E. coli*. The manufacturer's instructions were followed to prepare the culture medium and 25 g of LB powder was added to 1 litre of distilled water and mixed. The mixture was autoclaved for 15 minutes at 121°C to sterilise the medium. The medium was allowed to cool to room temperature and 100 µg/ml of ampicillin was added and thoroughly mixed. The prepared LB culture medium was stored at 4°C.

2.3.3.6.1.2. LB agar plates

Sigma-Aldrich's Lennox LB broth powder medium with agar (Sigma-Aldrich, Missouri, USA) was used to prepare petri dishes containing LB agar. The manufacturer's instructions were followed and 35 g of LB medium powder was added to 1 litre of distilled water. The total volume of LB agar medium to be made was adjusted by the number of LB agar petri dishes required for the spread plating (Figure 2.5) step of the cloning procedure. The medium was sterilised by following the standard autoclaving parameters of 15 minutes at 121°C. Ampicillin was added once the medium reached 55°C at a final concentration of 50 µg/ml in the InsTAclone™ PCR Cloning procedure and 100 µg/ml in the CloneJET™ PCR Cloning procedure and was gently mixed by swirling to avoid the introduction of bubbles. Approximately 25 ml of medium was poured into the required number of petri dishes and

allowed to solidify. Once the agar solidified, the plate was inverted, tightly sealed and stored at 4°C.

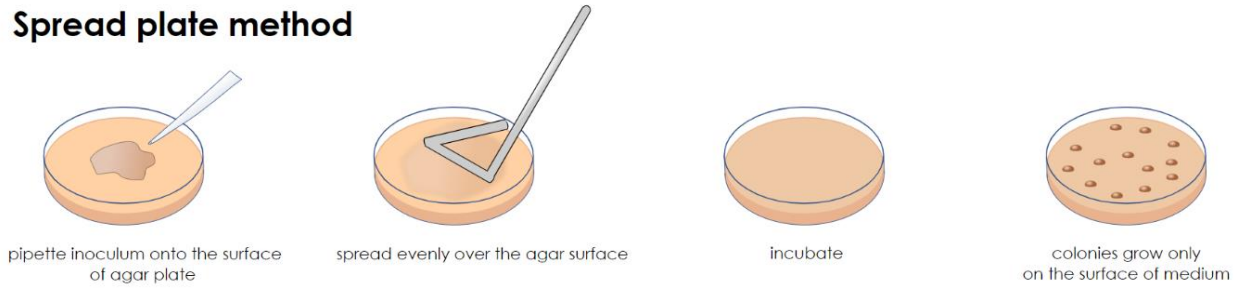

Figure 2.5: Diagram showing the spread plating method. Bacterial cells are added onto the surface of a solidified LB agar plate, a sterile plastic hook known as a hockey stick is used to evenly spread bacterial cells across the surface. The plate is inverted and incubated at 37°C to allow for the growth of bacterial cells. (Adapted from source: Macedo, 2016. [Online]: https://commons.wikimedia.org/wiki/File:Plating_methods.svg).

2.3.3.6.1.3. Ampicillin

An ampicillin stock solution (50 mg/ml) was prepared by dissolving 2.5 g ampicillin sodium salt (Tocris Bioscience, Bristol, UK) in 50 ml of deionized water. The solution was filter sterilized with a 0.22 µm syringe filter and stored in 1 ml aliquots at -20°C.

2.3.3.6.2. InsTAclone™ PCR Cloning Kit

Cloning Principle

This kit uses a TA cloning technique for direct cloning of products with 3'-dA overhangs that have been generated by the terminal transferase activity of *Taq* DNA polymerase during PCR amplification. The manufacturer's instructions were followed to ligate ISLA products into pTZ57R/T, a linearized cloning vector with a single 3'-ddT overhang at each end. The 3'-ddT overhangs at the ends of the vector promote cloning and prevent recircularisation of the vector during ligation, therefore 90% of recombinant clones contain the vector with an insert and have low background. Recombinant clones are selected using a process known as blue/white colony selection.

In brief blue/white colony selection works by manipulating the *lacZ* gene which is naturally found in bacterial cells. The *lacZ* gene encodes for β-galactosidase, an enzyme responsible for breaking lactose down into glucose and galactose. This enzymatic action is induced in

the presence of lactose but can also be induced by a lactose analogue known as isopropyl β-D-1-thiogalactopyranoside (IPTG) which is used in molecular cloning. In this molecular cloning technique, IPTG is used together with a dye-linked substrate, 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside, known as X-gal which results in the production of galactose and an insoluble blue pigment. The multiple cloning site (MCS) as shown in Figure 2.6, is contained within a partially deleted *lacZ* gene in pTZ57R/T which produces non-functional β-galactosidase. When pTZ57R/T is transformed into a bacterial cell with the lacZΔM15 mutation that encodes for the deleted section of the *lacZ* gene, the process known as alpha complementation occurs and functional β-galactosidase is produced. Successful ligation of the DNA fragment of interest into the pTZ57R/T MCS results in the interruption of the *lacZ* gene and the production of non-functional β-galactosidase which yields white-creamy coloured bacterial colonies. Unsuccessful ligation of the fragment into pTZ57R/T results in the production of blue coloured bacterial colonies as a result of functional β-galactosidase enzymatic activity in the presence of X-gal.

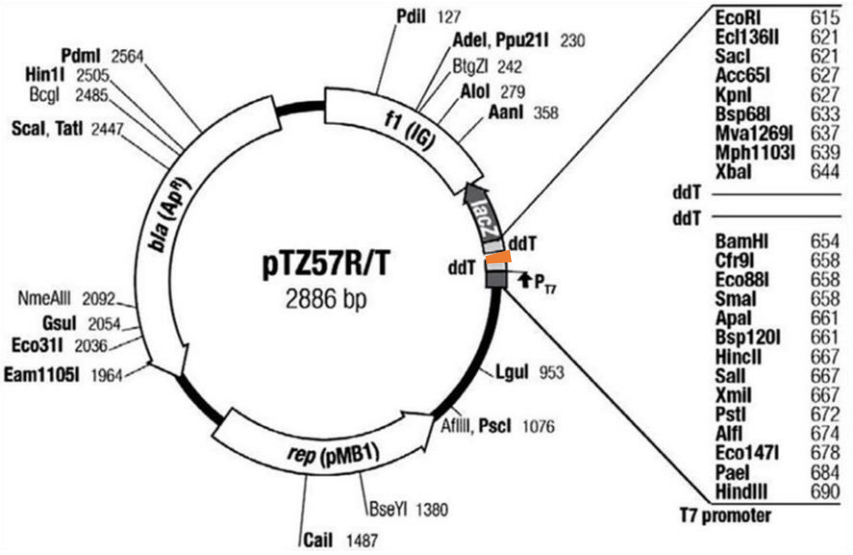

Figure 2.6: Map of the pTZ57R/T cloning vector from the InsTAclone™ PCR Cloning Kit. The MCS is shown in orange on the vector map. (Source: Thermo Scientific InsTAclone™ PCR Cloning Kit Manual, 2016).

Two control experiments were done in parallel. The positive control experiment used the Control PCR Fragment (0.52 pmol ends) provided in the kit. Nuclease-free water was used in the negative control experiment. The purpose of including the positive control was to ensure that the cloning procedure was optimal and the negative control was included to ensure that kit reagents were not contaminated.

2.3.3.6.2.1. Ligation of ISLA products into pTZ57R/T

The volume of PCR product added into the ligation reaction was calculated based on the information provided in the manufacturer's product information sheet, which stipulated that 0.17 ng per base pair be added into a reaction. The equation that was used to calculate the PCR product volume was $\frac{0.17 \times \text{size of fragment to be ligated}}{\text{ng}/\mu\text{l}}$. According to the manufacturer's instructions, the optimal ratio of insert to vector is 3:1. The following reagents from Thermo Scientific's InsTAclone™ PCR Cloning Kit were added into a 1.5 ml tube for the ligation reaction; 3 µl of vector pTZ57R/T (0.17 pmol ends), 6 µl of 5X Ligation Buffer and 1 µl of T4 DNA Ligase. The volume of ISLA product added was calculated using the equation above and 4 µl of the Control PCR Fragment (0.52 pmol ends) was added to the positive control reaction. Nuclease-free water was added to each reaction to bring the final volume to 30 µl. The ligation mixtures were incubated at room temperature for 1 hour.

2.3.3.6.2.2. Bacterial transformation of the recombinant pTZ57R/T

Invitrogen's One Shot™ TOP10 Chemically Competent *E. coli* (Invitrogen, California, USA) was in the bacterial transformation step. A tube containing 50 µl of the chemically competent *E. coli* was thawed on ice and 2.5 µl of the ligation mixture containing the ISLA product was added into the tube and gently mixed by flicking. For the positive control, 1 µl of the ligation mixture was added to 50 µl of chemically competent bacterial cells and gently mixed by flicking. The tubes were incubated on ice for 30 minutes followed by a heat shocking step for 30 seconds at 42°C to allow for the uptake of the plasmid vector into the bacterial cell. The tubes were immediately placed on ice following the heat shocking step for 2 minutes. Super Optimal broth with Catabolite repression (S.O.C.) Medium (Sigma-Aldrich, Missouri, USA) was prewarmed to 37°C and 250 µl was aseptically added to each tube containing the transformed *E. coli*. The tubes were incubated at 37°C in the Labcon Shaking Incubator 3081U (Labcon, California, USA) for 1 hour at 225 RPM. LB agar plates were removed from 4°C storage and 50 µg/ml of X-Gal ready-to-use Solution, 20 mM of IPTG ready-to-use Solution and 100 µg/ml of ampicillin were added to room temperature plates and spread across the surface using the spread plating technique. Three different volumes of the transformed bacterial solutions were selected for the spread plating step; 50 µl, 100 µl and 150 µl were used for each transformation. Each volume was spread across the surface of an LB agar plate that was prewarmed to 37°C for 30 minutes and was left to dry for between 20 to 30 minutes. The plates were inverted and incubated at 37°C overnight for a maximum

of 16 hours to avoid overgrowth and the formation of satellite colonies which are small colonies that do not contain the plasmid vector.

2.3.3.6.2.3. Colony Screening

The incubated plates were viewed for the presence of colonies. Blue and white-creamy coloured colonies were observed on the surface of the LB agar plates inoculated with the ISLA recombinant clones. The blue/white colony selection principle was explained in section 2.3.3.6.2. where blue colonies indicate that the colony does not contain the ISLA fragment and white-creamy colonies suggest that ligation was successful and these colonies were chosen for further analysis. A lawn of only blue colonies were present on the LB agar plates inoculated with bacteria transformed with the negative control. The LB agar plate inoculated with the positive control contained mostly white-creamy coloured colonies present on the surface apart from a few blue colonies that occurred as a result of unsuccessful ligation with the Control PCR Fragment.

2.3.3.6.2.4. Bacterial culture of the picked colony in LB broth

The chosen white-creamy colonies were carefully picked from the surface of the LB agar with a P10 pipette tip without piercing the agar. The P10 pipette tip used to pick the colony was carefully transferred into a 15 ml Falcon™ Round-Bottom Polypropylene Tube containing 5 ml of LB broth prewarmed to 37°C. These tubes were incubated at 37°C at a 60° angle in the Labcon Shaking Incubator rotating at 225 RPM for 12 to 16 hours.

2.3.3.6.2.5. Plasmid DNA purification of the overnight bacterial culture with the GeneJET Plasmid Miniprep Kit

The bacteria from the overnight culture were harvested by centrifugation at 4000 RPM for 5 minutes at room temperature. The bacterial cells formed a pellet at the base of the tube and the supernatant was discarded. Reagents from Thermo Scientific's GeneJET Plasmid Miniprep Kit (Thermo Scientific, Massachusetts, USA) were used to isolate the plasmid DNA from the bacterial cells. The cell pellet was resuspended in 250 µl of the Resuspension Solution and mixed by vortexing. The bacterial cells were lysed by adding 250 µl of the Lysis Solution and mixed by inverting the tube six times, after which 350 µl of the Neutralisation Solution was added and was mixed by inversion. The tube was centrifuged at maximum speed for 5 minutes. The supernatant was transferred into a Thermo Scientific GeneJET Spin Column, provided in the GeneJET Plasmid Miniprep Kit and centrifuged for 1 minute

at maximum speed to bind the DNA to the column filter. The bound DNA was washed by adding 500 µl of the Wash Solution, centrifuged for 1 minute and the flow-through was discarded. This step was repeated for a second wash. The empty column was centrifuged for 1 minute at maximum speed to remove any remaining Wash Solution. The filter column containing the bound plasmid DNA was transferred into a new 1.5 ml tube and 50 µl of Elution Buffer prewarmed to 70°C was added to the column and incubated for 2 minutes at room temperature. The tube was centrifuged for 2 minutes at maximum speed to collect the purified plasmid DNA and was visualised by separation through gel electrophoresis and the vector plasmids containing the fragment of interest, based on the base pair length of the fragment, were used for Sanger sequencing (2.3.3.6.4).

2.3.3.6.3. CloneJET™ PCR Cloning Kit

Cloning principle

A blunt-end cloning technique is used to ligate PCR products into the pJET1.2/blunt cloning vector provided in the CloneJET™ PCR Cloning Kit. The linearized cloning vector, pJET1.2/blunt shown in Figure 2.7, contains a lethal gene that is disrupted when a product is inserted into the cloning site which means that only cells with recombinant plasmids (vector containing a ligated insert) are able to propagate. Recircularised pJET1.2/blunt vector molecules that do not contain an insert will express a lethal restriction enzyme which will kill the host *E. coli* after transformation has occurred. This is a form of positive selection which accelerates the process of colony screening and eliminates additional costs required for blue/white colony selection. Phosphorylated (sticky-end) or non-phosphorylated (blunt-end) products can be ligated into pJET1.2/blunt. Products with overhangs will be processed with the sticky-end cloning protocol which removes phosphorylated ends and produces blunt-ended products prior to the ligation reaction. Recombinant clones are produced 99% of the time with this kit and all available *E. coli* strains can be directly transformed with the ligation product. The protocol described in the manufacturer's protocol was followed with some adjustments.

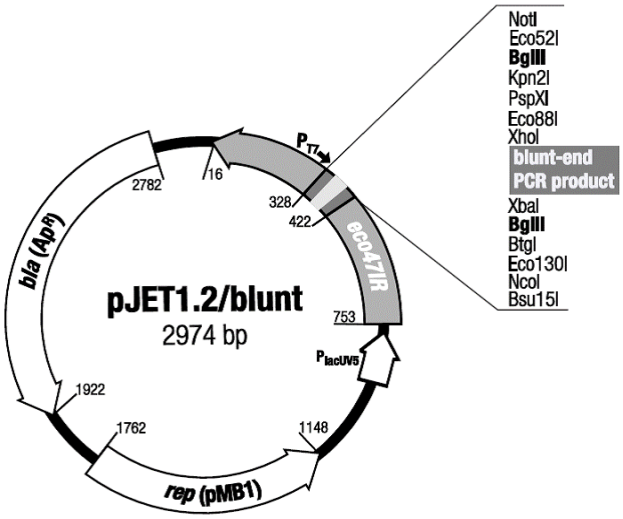

Figure 2.7: Map of the pJET1.2/blunt vector from the CloneJET™ PCR Cloning Kit. (Source: Thermo Scientific CloneJET™ PCR Cloning Kit Manual, 2019).

A control experiment was done in parallel following the manufacturer's instructions using the Control PCR Product (24 ng/µl) provided in the kit.

2.3.3.6.3.1. Ligation of ISLA products into pJET1.2/blunt

RANGER mix uses a *Taq* polymerase enzyme which generates products with 3'-dA overhangs, the sticky-end protocol was therefore followed with the purified ISLA PCR products. The overhangs were removed prior to ligation into pJET1.2/blunt in a reaction known as the blunting reaction. This reaction used 10 µl of 2X Reaction Buffer and 1 µl of DNA Blunting Enzyme from Thermo Scientific's CloneJET™ PCR Cloning Kit. Lastly, 7.5 µl of PCR product was added. For the positive control reaction, 2 µl of Control PCR Product (24 ng/µl) was added and 5 µl of nuclease-free water. The reactions were incubated at 70°C for 5 minutes and briefly chilled on ice.

The ISLA product was inserted into the vector in a reaction known as the ligation reaction and comprised of 0.5 µl of the pJET1.2/blunt Cloning Vector (50 ng/µl) and 1 µl of T4 DNA Ligase which was directly added to the blunting reaction. One microlitre of pJET1.2/blunt Cloning Vector and 1 µl of T4 DNA Ligase were added to the positive control reaction and both reactions were incubated at room temperature for 30 minutes.

2.3.3.6.3.2. Bacterial transformation of the recombinant pJET1.2/blunt

Invitrogen's One Shot™ TOP10 Chemically Competent *E. coli* was thawed on ice. Four microliters of the ligation reaction was combined with approximately 16 µl of the thawed chemically competent *E. coli* and incubated on ice for 30 minutes. Following the incubation, the *E. coli* cells were heat-shocked for 30 seconds at 42°C to allow for the uptake of the recombinant plasmid vector and immediately cooled on ice for 2 minutes. S.O.C. Medium was prewarmed to 37°C, 400 µl was added to the bacterial cells and incubated for 1 hour at 250 RPM in the Labcon Shaking Incubator at 37°C. After the 1-hour incubation, 100 µl of the transformed *E. coli* was spread across the surface of a LB agar plate prewarmed to 37°C using the spread plating technique. The remaining transformed *E. coli* was centrifuged at 3000 RPM for 3 minutes, 200 µl of supernatant was removed and the cell pellet was gently resuspended by pipetting and spread across a second prewarmed LB agar plate. The plates were left to dry for 20 to 30 minutes, inverted and incubated overnight for a maximum of 16 hours at 37°C.

2.3.3.6.3.3. Colony screening

The LB agar plates were viewed for the presence of white-creamy colonies following the overnight incubation. As the vector in this kit contains a lethal gene, all colonies present on the LB agar plate contain an insert.

2.3.3.6.3.4. Colony PCR with a high-fidelity enzyme

Single colonies present on the LB agar plates were carefully picked using a P10 pipette tip without piercing the agar. The P10 tip with a single colony was swirled in a 0.2 ml PCR tube containing the following PCR reagents for a confirmatory colony PCR; 10 µl of 2X RANGER Mix, 9.2 µl of nuclease-free water, 0.4 µl of pJET1.2 Forward Sequencing Primer (5'- CGA CTC ACT ATA GGG AGA GCG GC -3') and 0.4 µl of pJET1.2 Reverse Sequencing Primer (5'- AAG AAC ATC GAT TTT CCA TGG CAG -3') each at 10 µM.

The cycling parameters for the PCR step included an initial denaturation at 95°C for 3 minutes, 25 cycles cycling through 94°C for 30 seconds, 60°C for 30 seconds, 72°C for 1 minute.

2.3.3.6.3.5. Visualisation of the colony PCR products

The results of the colony PCR products were visualised through separation by gel electrophoresis and exposure to ultraviolet light as described in section 2.3.2.3.3.2 to determine whether ISLA products successfully ligated into pJET1.2/blunt. The base pair length of the ISLA fragment inserted into the vector is known and therefore, the expected size of the fragment present on the gel can be pre-calculated. Based on the information shown in Figure 2.8, the expected fragment size can be calculated by adding 118 base pairs to the inserted ISLA fragment length. The colony PCR products meeting this criterion were directly sequenced by Sanger sequencing.

Figure 2.8: DNA sequence preceding the MCS of pJET1.2/blunt and indicating the binding sites of the sequencing primers. The red arrows indicate how the 118 base pair length that is added to determine the expected size of a successfully ligated fragment is calculated. (Adapted from Source: Fermentas. [Online]: http://www.bioinfo.pte.hu/f2/pict_f2/pJETmap.pdf).

2.3.3.6.4. Sanger sequencing of the purified products from InsTAclone™ and CloneJET™ PCR Cloning Kits

The GeneJET purified products and the colony PCR products from the respective cloning kits were diluted to 20 ng/µl in nuclease-free water and prepared for sequencing with reagents from the BigDye™ Terminator v1.1 Cycle Sequencing Kit. The reagent volumes used for the preparation of the sequencing reaction was the same as described in 2.3.3.5.

The purified products cloned with the InsTAclone™ PCR Cloning Kit were sequenced using 1 µl of 5 µM M13/pUC Forward (5'- CCC AGT CAC GAC GTT GTA AAA CG -3') and 1 µl of 5 µM M13/pUC Reverse (5'- AGC GGA TAA CAA TTT CAC ACA GG -3') in two separate sequencing reactions. The thermocycling conditions used were as follows; 96°C for 0.10 seconds, 50°C for 0.05 seconds and 60°C for 4 minutes cycling through 25 cycles.

The selected colony PCR products generated through cloning with the CloneJET™ PCR Cloning Kit were sequenced using 1 µl of the pJET sequencing primers, pJET1.2 Forward

Sequencing Primer and pJET1.2 Reverse Sequencing Primer, at a concentration of 3.2 µM and in two separate sequencing reactions. The thermocycling parameters that were used included 25 cycles cycling through 96°C for 0.10 seconds, 60°C for 0.05 seconds and 60°C for 4 minutes.

Plates containing these reactions were sent to CAF at Stellenbosch University for sequencing clean-up and DNA sequencing.

2.3.3.6.5. Analysis of Sanger sequencing

NCBI's Nucleotide BLAST® (blastn) algorithm and program was used to align the raw Sanger sequences generated by direct sequencing (2.3.3.5.) or cloning followed by sequencing (2.3.3.6.4), to the nucleotide sequences present in the GenBank database (NCBI Resource Coordinators, 2016). The Sanger sequencing results were further aligned to the South African subtype C reference sequence Ref.C.ZA.04.04ZASK146.AY772699 (HIV Consensus C sequence) and/or the HXB2 HIV subtype B reference genome (GenBank: K03455.1) to analyse, generate consensus sequences and identify the sites of integration and the adjacent human genomic sequence using Geneious R11.1 (Biomatters Ltd, Auckland, New Zealand).

2.3.4. Oxford Nanopore Technologies– GridION sequencing

2.3.4.1. Rationale

The GridION sequencing device from ONT was investigated due to the several advantages this technology has to offer. The real-time sequencing technology was used to determine whether integration sites and provisionally intact NFL HIV-1 proviral sequences were recoverable and comparable to other sequencing platforms. The overall methodological approach is shown in Figure 2.9. Two different ONT protocols were investigated for these purposes. ONT's Premium whole genome amplification protocol was utilised in an attempt to identify HIV-1 integration sites from MDA amplified single genome NFL or p6-PR-RT pre-nested products and ONT's Amplicons by Ligation protocol was used to sequence provisionally intact NFL HIV-1 proviruses.

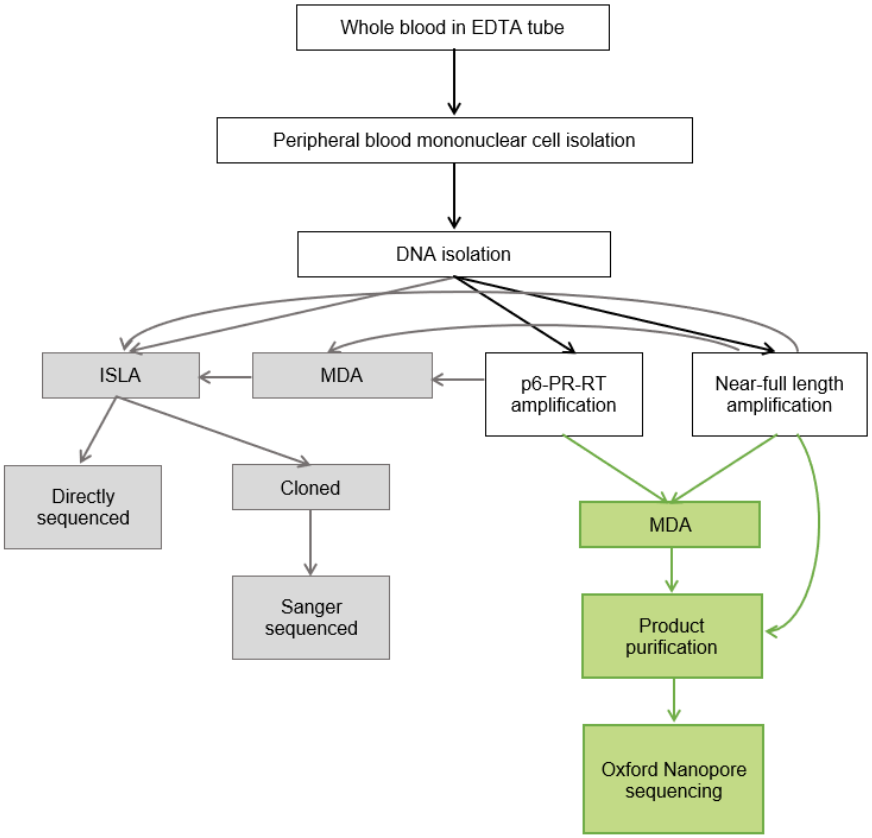

Figure 2.9: Overview of the methods used for ONT sequencing to recover HIV-1 integration sites and provisionally intact NFL HIV-1 proviruses.

2.3.4.2. Steps used in both ONT sequencing protocols

Step 1: *Determining the number of nanopores on the flow cell sensor array available for sequencing*

As part of ONT's quality control (QC), a minimum number of active nanopores are required to ensure successful sequencing which is determined by the type of flow cell that is used. ONT's FLO-MIN106D flow cells, depicted in Figure 2.10, were used for all sequencing experiments conducted on the GridION sequencing device. A flow cell was inserted into one of the five ports available on the device and the number of active nanopores was determined prior to sequencing initiation. The operating software that ONT uses to drive their sequencing devices, MinKNOW, was used to perform the Platform QC/flow cell check test to determine the number of available nanopores. The minimum number of active nanopores was limited to ≥800 of the 2048 nanopores present on the sensor array.

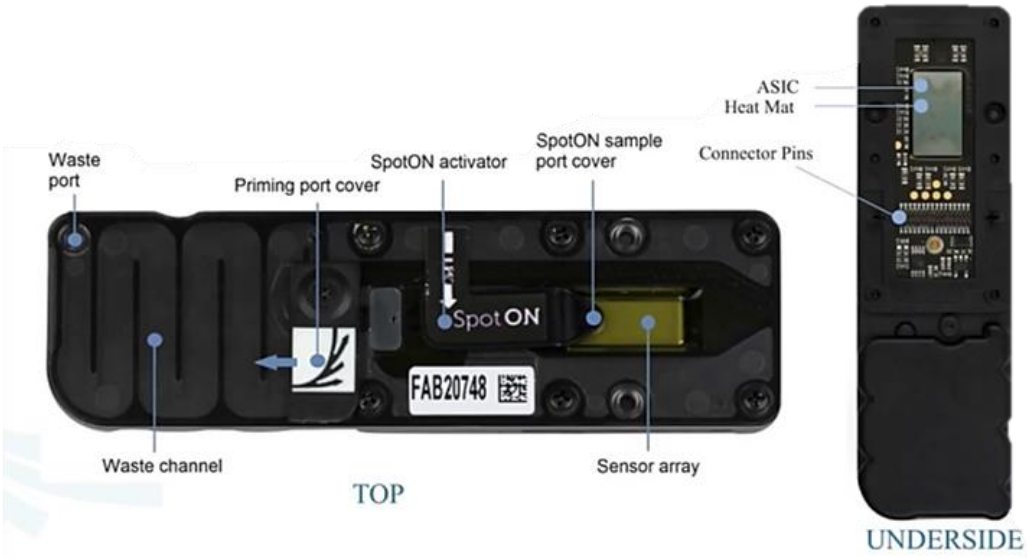

Figure 2.10: The components of the FLO-MIN106D flow cell used for sequencing with ONT's GridION. (Source: Oxford Nanopore Technologies, 2020. [Online]: https://www.youtube.com/watch?v=zC6lAtzqi_k).

Step 2: *Repairing and preparing the ends of DNA for sequencing adaptor attachment*

The ends of the DNA fragments were repaired to increase the library yield to generate more data and a higher percentage of coverage. The ends of the DNA fragments were repaired with the NEBNext® FFPE DNA Repair Mix (New England Biolabs, Massachusetts, USA) that contains enzymes that repair the deamination of cytosine to uracil, nicks and gaps, oxidised bases and the blocked 3'ends of DNA. The ends of the DNA were prepared for the attachment of the sequencing adaptors with the NEBNext® Ultra™ II End Repair/dA-tailing Module (New England Biolabs, Massachusetts, USA) which processes the 3' and 5' ends of DNA to contain 3' dA-tailed ends and 5' phosphorylated ends.

The reparation and preparation of the ends of the DNA was carried out in a 0.2 ml PCR tube. The concentration of DNA added was dependent on the ONT sequencing protocol that was followed. The input concentrations were 1.5 µg of MDA amplified DNA for the Premium whole genome amplification protocol and 1 µg of nested NFL amplicons for the Amplicons by Ligation protocol. For both protocols, the DNA was diluted to a total volume of 48 µl in nuclease-free water. Reagents from the two NEB kits were added to the 0.2 ml PCR tube in the following order; 3.5 µl of NEBNext® FFPE DNA Repair Buffer, 2 µl of NEBNext® FFPE DNA Repair Mix, 3.5 µl of NEBNext® Ultra™ II End Repair Reaction Buffer and lastly 3 µl of NEBNext® Ultra™ II End Repair Enzyme Mix. The reaction was gently mixed by flicking and incubated in a thermocycler for 5 minutes at 20°C to repair and prepare the ends of the DNA and 5 minutes at 65°C to inactivate the enzymes.

Step 3: *Attaching the sequencing adaptors to the ends of the prepared DNA fragments*

ONT's 1D sequencing adaptors contain a leader sequence and a motor protein. The leader sequence assists the DNA in locating and binding to the nanopores on the sensor array which facilitates in initiating sequencing. The motor protein assists in unwinding the double-stranded DNA fragment and slows the rate at which the single-stranded DNA is sequenced by the nanopore.

The end-repaired and prepared DNA from *Step 2* was either used directly (Premium whole genome amplification) or purified (Amplicons by Ligation) prior to use in this step depending on the ONT protocol that was followed. Sixty microliters of the product directly from *Step 2* or the purified end prepared product (2.3.4.4.2.) was added into a 1.5 ml Eppendorf® LoBind DNA tube. In addition to this, 25 µl of Ligation Buffer (LNB) and 5 µl Adaptor Mix (AMX) from ONT's Ligation Sequencing Kit (SQK-LSK109) were added followed by 10 µl of Quick T4 DNA Ligase from NEB's NEBNext® Quick Ligation Module (New England Biolabs, Massachusetts, USA). The tube was incubated at room temperature for 10 minutes to attach the 1D sequencing adaptors to the ends of the DNA fragment.

Step 4: *Priming the flow cell for sequencing*

All ONT flow cells are shipped with a storage buffer that covers the sensor array and protects the nanopores. The storage buffer is replaced with a priming solution prior to sequencing. The priming solution is comprised of a tether and a buffer. The tether plays an essential role in concentrating the DNA at the surface of the membrane and streamlines the sequencing while the buffer is a mixture of running buffer and nuclease-free water which provides optimal conditions and the fuel required for successful sequencing.

The priming port cover is opened and 20 – 30 µl of the storage buffer is carefully aspirated using a P1000 pipette from the port to remove the air bubble and ensure that bubbles are not introduced into the sensor array in later steps which will damage and inactivate the nanopores. The priming solution was prepared by adding 30 µl of Flush Tether (FLT) directly to a tube of Flush Buffer (FB) from ONT's Flow Cell Priming Kit (EXP-FLP002). The priming solution was mixed by pipetting and 800 µl was slowly added into the priming port without introducing bubbles thereby displacing the storage buffer from the sensor array into the waste channel. After adding the priming solution, the flow cell is incubated for 5 minutes at

room temperature followed by slowly adding a further 200 µl of the priming solution into the priming port after opening the SpotON port cover and avoiding the introduction of bubbles.

Step 5: *Preparing the library for loading into the SpotON port of the flow cell*

The prepared DNA library (DNA with adaptors attached) is combined with loading beads that function to bring the DNA closer to the pores at the start of sequencing and also result in high-yield and high-speed sequencing.

The DNA library was purified prior to sequencing and 12 µl was added into an Eppendorf® LoBind DNA tube. The Loading beads (LB) from ONT's Ligation Sequencing Kit were vortexed to form a homogenous mixture and 25.5 µl was added to the purified DNA library. Lastly, 37.5 µl of Sequencing Buffer (SQB) from ONT's Ligation Sequencing Kit was added into the tube. The mixture was gently mixed before it was added into the SpotON port on the flow cell. After priming the flow cell (Step 4), 75 µl of the prepared library was loaded into the SpotON port in a dropwise fashion ensuring that each drop successfully flowed into the port before closing the SpotON port cover.

Step 6: *Initiating the sequencing run with ONT's MinKNOW software*

The MinKNOW software is also responsible for controlling the sequencing device and includes the ability to select the running parameters, sample identification and tracking and also ensuring that the platform chemistry is optimal for successful sequencing. Furthermore, MinKNOW ensures the following core functions are carried out; data acquisition, real-time analysis and feedback, local basecalling and data streaming (Oxford Nanopore Technologies, Nanoporetech, 2020). The running parameters are selected by the operator and allow for the type of sequencing to be selected which is determined by the sequencing kit used to prepare the sample, the preferred method for basecalling, the duration of the sequencing run and the preferred output of the sequencing data in FAST5 and/or FASTQ files.

The sequencing runs were initiated after Steps 1 to 5 were completed and the flow cell stabilised at 34°C. The selected samples were sequenced under the following run parameters; LSK-SQK109 Kit for DNA sequencing, fast-basecalling method and both FAST5 and FASTQ files were selected as outputs for the sequencing data. The duration of

the run was left as standard and the run was stopped when sufficient coverage was obtained.

2.3.4.3. ONT's Premium whole genome amplification protocol

ONT's Premium whole genome amplification protocol was utilised in an attempt to recover the 8E5 HIV-1 integration site from NFL or p6-PR-RT pre-nested single genome products that were MDA amplified. An overview of the Premium whole genome amplification protocol is highlighted in Figure 2.11.

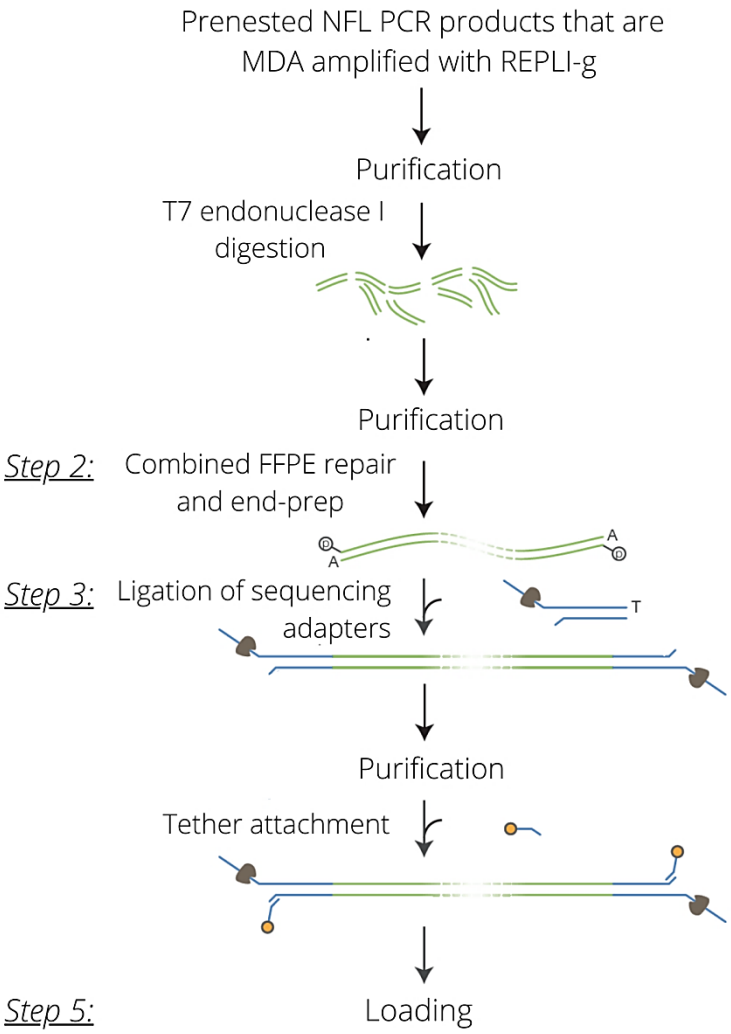

Figure 2.11: Flow diagram showing the steps involved in ONT's Premium whole genome amplification protocol (version: WAL_ 9070_v109_revK_14Aug2019). (Adapted from source: Oxford Nanopore Technologies' Premium whole genome amplification protocol, 2019).

2.3.4.3.1. Purification of MDA amplified PCR products prior to ONT sequencing

Successful sequencing with ONT is dependent on good DNA yield and high purity. The products, therefore, need to be free of impurities created by excessive amounts of salts from buffers, unincorporated dNTPs, primers, primer dimers and other contaminants, as these may result in the blockage of the nanopores leading to unsuccessful sequencing. Several methods were investigated to determine the best method of purification.

2.3.4.3.1.1. AMPure XP paramagnetic bead purification of PCR products

ONT recommends purifying products for sequencing with Agencourt AMPure XP PCR Purification beads (referred to as AMPure XP beads in the following sections) (Beckman Coulter Life Sciences, California, USA) which uses a buffer to bind DNA fragments to paramagnetic beads, removes contaminants and results in the best recovery of the fragment of interest.

The manufacturer's instructions were followed and 50 µl of the MDA amplified sample was transferred into a 1.5 ml Eppendorf® LoBind DNA tube (Sigma-Aldrich, Missouri, USA). The AMPure XP beads were equilibrated to room temperature. The beads were vortexed to form a homogenous mixture before being added to the sample. The ratio of sample to beads that was used was 1:1.8, therefore 90 µl of resuspended AMPure XP beads were added to the sample. The mixture was gently mixed by pipetting and incubated at room temperature in the Labnet Vortemp 56 shaking incubator (Labnet International Incorporated, New Jersey, USA) for 5 minutes at 90 RPM. The sample was briefly centrifuged and placed onto the DynaMag™-2 Magnet's magnetic rack (Invitrogen, California, USA) to pellet the beads. The supernatant was removed and the sample was washed with 500 µl of freshly prepared 70% ethanol without disturbing the pellet. The ethanol was carefully removed and the ethanol wash step was repeated. The tube was briefly centrifuged after removing the supernatant from the second wash step. The tube was placed onto the magnetic rack to remove any residual supernatant and the lid of the tube was opened to dry the pellet for 1 minute. The tube was removed from the magnetic rack and the pellet was resuspended in 50 µl of 5 mM Tris-HCl. The resuspended mixture was incubated at room temperature for 2 minutes. The beads were pelleted after placing the tube onto the magnetic rack and the eluate containing the purified nucleic acid was transferred into a new 1.5 ml Eppendorf® LoBind DNA tube. The purity and concentration of the DNA was measured using the NanoDrop™ 1000 Spectrophotometer.

2.3.4.3.1.2. Ethanol precipitation

Ethanol precipitation is a method that is commonly used to purify and/or concentrate DNA present in aqueous solutions by adding ethanol as an antisolvent. The basic principle is that molecular grade ethanol and a salt, usually sodium acetate, are added to the aqueous solution with ethanol being more soluble, displacing nucleic acid salt, which precipitates out of the solution. The nucleic acid salt separation is accelerated by centrifugation, with the precipitant forming a pellet on the outer side of the tube. The pellet is usually washed with cold 70% ethanol to dissolve residual salt, however, the concentration of the ethanol used is dependent on the precipitation method. The ethanol is removed without disturbing the pellet and the pellet is dried before being resuspended in an aqueous buffer such as nuclease-free water, 5 mM Tris-HCl or 1x TE buffer (pH 8.0) as nucleic acids are soluble in these solutions.

2.3.4.3.1.2.1. Qiagen's purification of MDA amplified products

The MDA amplified products were purified using Qiagen's Purification of DNA amplified using REPLI-g® Kits (version RG21 Nov-12). The manufacturer's instructions were followed and 50 µl of MDA amplified product was equilibrated to room temperature and transferred into a 1.5 ml tube. Following this, 150 µl of 96 – 100% molecular grade ethanol was added to the tube and gently mixed by flicking. The mixture was centrifuged at maximum speed for 2 minutes to pellet the DNA. The supernatant was carefully removed without disturbing the pellet and 100 µl of freshly prepared 80% ethanol was carefully added without resuspending the pellet. The mixture was centrifuged at maximum speed for 2 minutes. The supernatant was discarded and the tube was briefly centrifuged to collect and remove any residual supernatant at the bottom of the tube. The precipitated nucleic acid was incubated at room temperature for 5 minutes to evaporate any remaining ethanol. The precipitated nucleic acid was resuspended in 50 µl of 1x TE buffer (pH 8.0) and was measured with a Qubit® 2.0 Fluorometer (Invitrogen, California, USA) and the NanoDrop™ 1000 Spectrophotometer.

2.3.4.3.1.2.2. MRC Holland ethanol precipitation

The ethanol precipitation protocol published by MRC Holland was also investigated. Fifty microliters of MDA amplified product, 5 µl of 3 M sodium acetate (pH 5.2) and 150 µl of ≥95% molecular grade ethanol was added to a 1.5 ml tube and mixed by pipetting. The reaction mixture was incubated on ice for 15 minutes. Following the incubation, the reaction

mixture was centrifuged at maximum speed for 30 minutes at room temperature and the supernatant was carefully removed without disturbing the pellet. The pellet was washed with 1 ml of freshly prepared 70% ethanol and centrifuged at maximum speed for 15 minutes. The supernatant was discarded and the pellet was briefly centrifuged to remove any residual supernatant. The pellet was air-dried with an opened lid until the area around the pellet changed colour and the pellet was resuspended in 50 µl of 5 mM Tris-HCl. The concentration of the purified DNA was measured with the NanoDrop™ 1000 Spectrophotometer.

2.3.4.3.1.2.3. In-house ethanol precipitation protocol

The purification protocol was carried out in a 1.5 ml tube to which 50 µl of MDA amplified product, 5 µl of 125 mM EDTA, 5 µl 3 M sodium acetate (pH 5.2) and 125 µl 100% molecular grade ethanol was added. The reaction mixture was vortexed and incubated at room temperature for 15 minutes. The tube was centrifuged for 20 minutes at maximum speed. The supernatant was discarded without disturbing the pellet and 500 µl of freshly prepared 70% molecular grade ethanol was added without resuspending the pellet. The reaction was centrifuged for 10 minutes at maximum speed and the supernatant was carefully removed. The pellet was dried for 5 minutes or until no ethanol was visible in the tube. The pellet was resuspended in 50 µl of 5 mM Tris-HCl and the purified DNA was measured with a Qubit® 2.0 Fluorometer and the NanoDrop™ 1000 Spectrophotometer.

2.3.4.3.2. Removing the hyperbranched structures produced by MDA amplification

The purified MDA amplified products were treated with the T7 Endonuclease I enzyme (New England Biolabs, Massachusetts, USA) to resolve the hyperbranched structures generated by MDA amplification to allow for a higher average ONT Q score to be obtained when sequencing. Sequencing hyperbranched structures results in a rapid blockage of nanopores and therefore a decline in sequencing yield and the average Q score.

According to ONT's protocol, 1.5µg of purified MDA amplified DNA was added to a 0.2 ml PCR tube together with 3 µl of NEBuffer 2, 1.5 µl T7 Endonuclease I and lastly nuclease-free water to bring the final reaction volume to 30 µl. The reaction was incubated for 15 minutes at 37°C in a thermocycler.

A large decrease in the concentration of the product was observed following the T7 Endonuclease I digestion. This problem was also encountered by other ONT users and this

step was optimised by increasing the input concentration of the purified MDA amplified DNA to 3.5 µg as recommended on ONT's community dashboard. The remaining reagents were added in the volumes mentioned above and the nuclease-free water volume was adjusted to bring the final reaction volume to 30 µl. The reaction was incubated at 37°C for 10 minutes and 80°C for 5 minutes to inactivate the enzyme.

2.3.4.3.3. Purification of the T7 Endonuclease I digested products using AMPure XP beads

The T7 Endonuclease I digested products were purified prior to *Step 2* to remove the remaining buffer and T7 Endonuclease I enzyme. Several purification methods were investigated to determine which would give the best purity and DNA yield as a large decrease in the concentration of the MDA amplified product was seen following the T7 Endonuclease I digestion.

2.3.4.3.3.1. AMPure XP bead purification using a custom buffer system

The Premium whole genome amplification protocol from ONT recommended that T7 Endonuclease I digested products be purified with AMPure XP beads that have been resuspended in a custom buffer. The original AMPure XP bead buffer was therefore replaced with a custom buffer. The custom buffer comprised of 10 mM Tris-HCl, 1 mM EDTA (pH 8.0), 1.6 M NaCl, 10 mM PEG 8000 and 444 µl of nuclease-free water in a final volume of 996 µl. To replace the original buffer with the custom buffer, the AMPure XP beads were resuspended to form a homogenous mixture and 1 ml was transferred into a 1.5 ml Eppendorf® LoBind DNA tube. The beads were pelleted by placing the tube onto the magnetic rack and the original buffer was removed. The beads were resuspended and washed in 1 ml of nuclease-free water. The tube was returned to the magnetic rack and the supernatant was removed after the beads formed a pellet. The nuclease-free water wash step was repeated to remove any remaining original buffer. After removing the nuclease-free water from the second wash step, the tube was briefly centrifuged and placed onto the magnetic rack to remove any residual supernatant. The pellet was first resuspended in 200 µl of the custom buffer and then transferred into the tube containing the remaining custom buffer.

The 30 µl reaction mixture from the T7 Endonuclease I step was added to 20 µl of 1X TE buffer (pH 8.0) in a 1.5 ml Eppendorf® LoBind DNA tube. The custom AMPure XP bead

suspension was vortexed to form a homogenous mixture and 35 µl was added to the sample. The reaction was mixed by flicking and incubated on the HulaMixer (Thermo Fisher Scientific, Massachusetts, USA) for 20 minutes at room temperature to mix the reaction. The tube was briefly centrifuged and placed onto the magnetic rack to pellet the beads. The supernatant was removed and the pellet was washed with 200 µl of freshly prepared 80% ethanol without disturbing the pellet. The supernatant was discarded and the ethanol wash step was repeated. After removing the supernatant from the second wash, the tube was briefly centrifuged and placed back onto the magnetic rack to remove any residual supernatant. The pellet was dried for 30 seconds with an opened lid and the pellet was then resuspended in 49 µl of nuclease-free water. The resuspended mixture was incubated for 1 minute at 50°C in the Labnet AccuBlock™ Digital Dry Bath and was followed by a 5-minute incubation at room temperature. The tube was placed onto the magnetic rack to pellet the beads and the eluate was transferred into a new 1.5 ml Eppendorf® LoBind DNA tube. The quantity and quality of the purified nucleic acid was measured with the NanoDrop™ 1000 Spectrophotometer.

2.3.4.3.3.2. AMPure XP bead purification

The purification of T7 Endonuclease I digested products was attempted using the AMPure XP bead protocol described in section 2.3.4.3.1.1.

The AMPure XP bead purification method was also attempted on a MDA amplified NFL pre-nested HIV-1 positive patient sample with a sample to bead ratio of 1:0.8. The purification results of the patient sample and both MDA amplified NFL and p6-PR-RT pre-nested products were compared in terms of purity and were also visualised by plate viewing (section 2.3.2.3.3.1.) and gel electrophoresis (section 2.3.2.3.3.2.).

2.3.4.3.3.3. MRC Holland ethanol precipitation

The MRC Holland ethanol precipitation method was identified as the best purification method for purifying MDA amplified products prior to sequencing (section 2.3.4.3.1.) and was therefore also investigated for the purification of T7 Endonuclease I digested products. The method described in section 2.3.4.3.1.2.2. was followed. An additional optimisation step to the method was investigated in this section where two different incubation conditions were simultaneously investigated to determine which would result in the best purity and DNA yield. Therefore, one reaction mixture was incubated on ice for 15 minutes as described

previously and a second reaction mixture was incubated at -20°C overnight. Following the two independent incubation steps, the reaction mixtures were processed as mentioned in section 2.3.4.3.1.2.2. After the pellets were resuspended in 50 µl of 5 mM Tris-HCl, the concentration of the purified DNA was measured and compared using the NanoDrop™ 1000 Spectrophotometer.

2.3.4.3.3.4. NucleoTraP®CR silica bead-based purification

The NucleoTraP®CR Kit from Macherey-Nagel was investigated as a method for purifying the T7 Endonuclease I digested products. The manufacturer's instructions for the direct purification of PCR products was followed. The reaction was carried out in a 2 ml tube to which 4 volumes of Buffer NT2 was added to 1 volume of MDA amplified sample. The NucleoTraP®CR Suspension, which contains activated silica beads, was vortexed to form a homogenous mixture and 10 µl was added to each 100 µl of reaction mixture. The mixture was incubated at room temperature for 10 minutes and vortexed every 2 minutes during the incubation step to bind the DNA to the silica beads. The reaction mixture was centrifuged for 30 seconds at 10 000 RCF to pellet the silica beads and the supernatant was discarded. The pellet was washed with 400 µl of Buffer NT2 and the mixture was briefly vortexed to resuspend the pellet. The tube was centrifuged at 10 000 RCF for 30 seconds and the supernatant was removed. The next wash step was carried out by adding 400 µl of the ethanolic Buffer NT3 and resuspending the pellet by vortexing. The tube was centrifuged for 30 seconds at 10 000 RCF and the supernatant was discarded. A wash with Buffer NT3 was repeated and after removing the supernatant, the pellet was centrifuged for a second time to remove any residual Buffer NT3. The pellet was dried at 37°C for 15 minutes using the Labnet AccuBlock™ Digital Dry Bath (Labnet International Incorporated, New Jersey, USA). The pellet was resuspended in 50 µl of 5 mM Tris-HCl. The mixture was incubated at room temperature for 15 minutes and vortexed every 2 minutes during the incubation step. The sample was centrifuged for 30 seconds at 10 000 RCF to pellet the silica beads and the purified DNA was transferred into a new 1.5 ml tube. The quality and quantity of the purified DNA was measured with the NanoDrop™ 1000 Spectrophotometer.

2.3.4.3.3.5. NucleoSpin® column-based purification

Macherey-Nagel's NucleoSpin® Gel and PCR Clean-up Kit (Macherey-Nagel, Düren, Germany) was also investigated and the manufacturer's instructions were followed to purify the T7 Endonuclease I digested products. The total volume of the sample to be purified was

determined and 1 volume of sample was mixed with 2 volumes of Buffer NTI from the NucleoSpin® Kit. A maximum of 700 µl of the sample and Buffer NTI mixture was added into a NucleoSpin® Gel and PCR Clean-up Column in a 2 ml collection tube. The column and collection tube was centrifuged for 30 seconds at 11 000 RCF to bind the DNA to the silica membrane in the column and the flow-through was discarded. This step was repeated if any sample and Buffer NTI mixture remained. The bound DNA was washed by adding 700 µl of Buffer NT3 to the column. The tube was centrifuged for 30 seconds at 11 000 RCF and the flow-through was discarded. The wash step with Buffer NT3 was repeated and the tube was centrifuged for 30 seconds at 11 000 RCF. The silica membrane was dried by a centrifugation step at 11 000 RCF for 1 minute to ensure that any remaining Buffer NT3 was removed. The column was placed into a new 1.5 ml tube. Buffer NE was prewarmed to 70°C and 20 µl was added directly to the silica membrane to elute the bound DNA. The column was incubated at room temperature for 1 minute and this was followed by a 1-minute centrifugation at 11 000 RCF to collect the eluate. The purified DNA was measured by the NanoDrop™ 1000 Spectrophotometer.

2.3.4.3.4. Purification of the DNA library using AMPure XP beads

The purification of the DNA library from the Premium whole genome amplification protocol was followed with an adjustment to the ratio of AMPure XP beads added and the incubation instrument. The ONT protocol used a sample to AMPure XP bead ratio of 1:0.4, however, this ratio was changed to 1:0.8 to ensure that the product of interest was present in the eluate. The 100 µl reaction mixture from the adaptor ligation step was transferred into a 1.5 ml Eppendorf® LoBind DNA tube and 80 µl of resuspended AMPure XP beads were added. The reaction was mixed by flicking and the tube was placed into the Labnet Vortemp 56 shaking incubator rotating at 90 RPM for 5 minutes at room temperature. The beads were pelleted by placing the tube onto the magnetic rack and the supernatant was removed. The pellet was resuspended and washed in 250 µl of the Short Fragment Buffer (SFB) from ONT's Ligation Sequencing Kit and returned to the magnetic rack. The supernatant was removed after a pellet formed and the SFB wash was repeated. After the second wash and removal of the supernatant, the tube was briefly centrifuged to remove any residual supernatant. The pellet was dried for 30 seconds and resuspended in 15 µl of Elution Buffer (EB) from ONT's Ligation Sequencing Kit. The reaction mixture was incubated at 37°C in the Labnet AccuBlock™ Digital Dry Bath for 10 minutes. The tube was placed onto the magnetic rack to pellet the beads and the eluate was removed and placed into a new 1.5 ml

Eppendorf® LoBind DNA tube. The concentration of the DNA library was measured with the Qubit® 2.0 Fluorometer.

2.3.4.4. ONT's Amplicons by Ligation protocol

Provisionally intact NFL HIV-1 proviral sequences were previously identified through the Illumina® MiSeq™ sequencing platform. ONT's GridION performance characteristics in terms of sequencing yield and read accuracy were compared to the Illumina® MiSeq™ using amplicons generated from the same single genome amplified pre-nested product. The original NFL nested product was depleted for Illumina® sequencing and the pre-nested NFL products were used to regenerate nested NFL amplicons. The Amplicons by Ligation protocol from ONT as indicated in Figure 2.12 was used to sequence the provisionally intact nested NFL amplicons on the GridION.

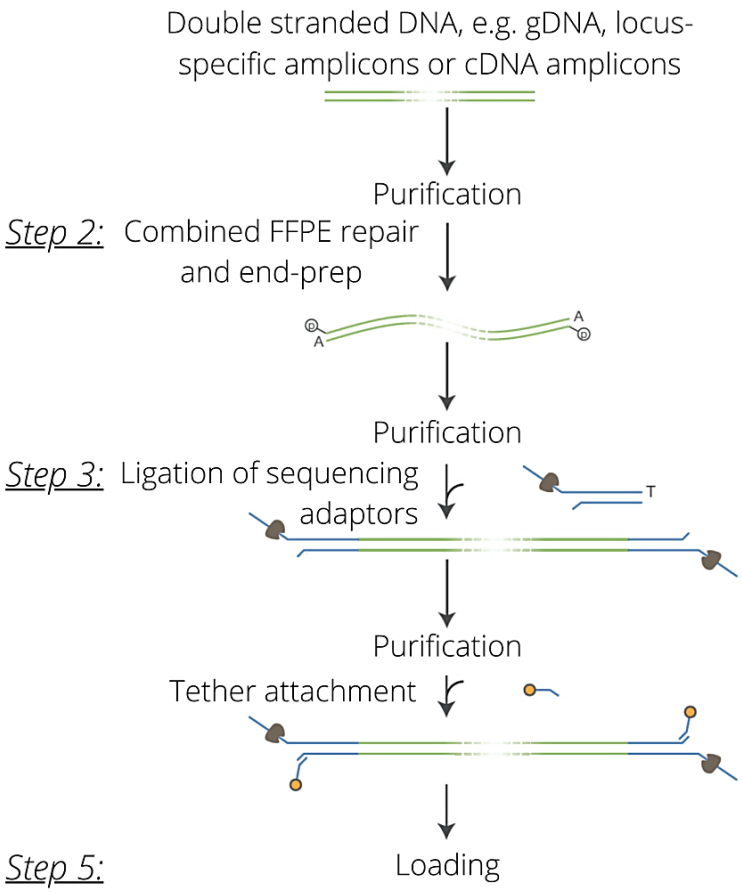

Figure 2.12: Flow diagram of the steps in ONT's Amplicons by Ligation protocol (version: ACDE_9064 _v109_revN_14Aug2019) (Adapted from source: Oxford Nanopore Technologies' Amplicons by Ligation, 2019).

2.3.4.4.1. Purification of provisionally intact NFL HIV-1 amplified PCR products prior to ONT sequencing

The nested NFL amplicons were purified using AMPure XP beads prior to ONT sequencing. Thirty microliters of the nested NFL product was added to a 1.5 ml Eppendorf® LoBind DNA tube. The purification steps were the same as described in section 2.3.4.3.1.1. with a few adjustments which will be highlighted in this section. A sample to beads ratio of 1:0.8 was used and the initial incubation conditions remained the same. After placing the tube containing the beads and the sample mixture onto the magnetic rack, the pellet was washed with 250 µl of freshly prepared 80% ethanol in both wash steps. The ethanol was removed after being in contact with the pellet for 30 seconds. The ethanol wash was repeated and after removing the ethanol from the second wash, the pellet was left to dry until the surface of pellet changed from glossy to matte in appearance and no ethanol was visible around the pellet. The pellet was resuspended in 50 µl of 5 mM Tris-HCl and the tube was incubated in the Labnet Vortemp 56 for 5 minutes at 37°C rotating at 90 RPM to increase the recovery and yield of the nucleic acid. The beads were pelleted on the magnetic rack and the eluate was placed into a new tube. The concentration and purity of the purified amplicon was measured with the NanoDrop™ 1000 Spectrophotometer.

2.3.4.4.2. Purification of the end-repaired and prepared DNA using AMPure XP beads

The DNA from ONT's DNA repair and end preparation step was purified with AMPure XP beads following the instructions in the Amplicons by Ligation protocol. The purification steps explained in section 2.3.4.3.1.1. was followed with a few adjustments aimed at increasing the recovery and purity of the prepared DNA. In brief, a 1:1 ratio of sample to AMPure XP beads was used. The initial incubation remained the same and both wash steps used 200 µl of freshly prepared 80% ethanol to wash the pellet. The ethanol was removed after being in contact with the pellet for 30 seconds. The pellet was dried until a change in appearance was observed and the pellet was resuspended in 60 µl of nuclease-free water. The resuspended reaction was incubated at room temperature for 2 minutes. The beads were pelleted on the magnetic rack and the eluate was removed and transferred into a new tube.

2.3.4.4.3. Purification of the DNA library using AMPure XP beads

The DNA library was also purified using AMPure XP beads as recommended by ONT and the steps described in the Amplicons by Ligation protocol were followed. In brief, a 1:0.4 ratio of sample to beads was used. The DNA library and bead mixture was incubated in the

Labnet Vortemp 56 shaking incubator at room temperature for 5 minutes at 90 RPM. The sample was briefly centrifuged and placed onto the magnetic rack to pellet the beads and remove the supernatant. The pelleted beads were resuspended and washed in 250 µl of the Long Fragment Buffer (LFB) from ONT's Ligation Sequencing Kit. The tube was returned to the magnetic rack to pellet the beads and remove the supernatant. The wash step was repeated for a second time using LFB, after the removal of the supernatant the tube was briefly centrifuged and returned to the magnetic rack to remove any residual LFB. The pellet was dried for 30 seconds and resuspended in 15 µl of EB. The reaction was incubated at 37°C in the Labnet AccuBlock™ Digital Dry Bath for 10 minutes. The beads were pelleted on the magnetic rack and the purified DNA library was transferred into a new tube.

2.3.4.5. Pipeline development for ONT analysis

A pipeline was developed for analysing the large amount of data that was generated by ONT sequencing by a collaborator affiliated with Stellenbosch University. In brief, the bases were recalled from the FAST5-pass files using the Guppy basecaller v3.3.3+fa743a6 (Oxford Nanopore Technologies, Oxford, UK). The read files were concatenated and aligned to the HIV Consensus C sequence using minimap2v2.17-r974-dirty (Li, 2018). The mapped reads were sorted and indexed using Samtools v1.10 (Li *et al*., 2009). Nanopolish v0.11.3 (Loman *et al*., 2015) was used to index the reads and map them back to original FAST5-pass files. Nanopolish was used in the "faster" mode and with the "fix homopolymers" feature on to call variants from the mapping. The maximum haplotypes considered was increased to 3000. An initial consensus was generated using Nanopolish's vcf2fasta tool. A second round of remapping was completed where only single nucleotide polymorphisms (SNP) were considered to allow more potential SNP haplotypes to be included in variable loops and to further polish the consensus.

Chapter 3

3 Results

3.1 Single genome amplification

3.1.1. Plate view of NFL and p6-PR-RT nested PCR amplicons

The near-full-length (NFL) and p6-PR-RT PCRs were performed as described in sections 2.3.2.3.1. and 2.3.2.3.2. and the nested amplicons from both PCRs were viewed using the plate viewing method described in section 2.3.2.3.3.1. The plate viewing method allowed for the estimation of the target dilution, a concentration at which a positive well would most likely represent a single genome in a well. A 2-fold and 3-fold serial dilution of 8E5 DNA was used and amplified with NFL and p6-PR-RT PCRs respectively, where each row on the plate represented a specific dilution ratio in the dilution series. According to Poisson's distribution, at a dilution where 4 wells out of 12 (30%) wells are positive, as detected by fluorescence of the EZ Vision™ DNA Dye, it is likely that the large majority of positive wells represent amplification from a single genome. Figure 3.1 shows the results of serially diluted 8E5 DNA amplified by the p6-PR-RT PCR and identifies the encircled 1:27 row as the target dilution. The method for identifying the target dilution for NFL was the same, however, the endpoint dilution was calculated based on the number of 8E5 HIV copies added to the reaction and the dilution target was identified as 15 copies per reaction.

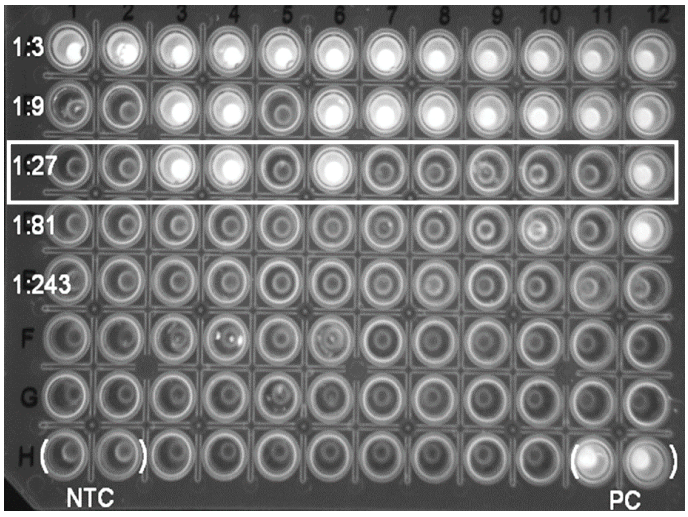

Figure 3.1: Plate viewing result of nested p6-PR-RT amplicons using serially diluted 8E5 DNA as a template to determine the target dilution. The respective dilution factors are indicated by *1:3*, *1:9*, *1:27*, *1:81* and *1:243. NTC*, non-template control, *PC,* positive control.

3.1.2. Gel electrophoresis of NFL and p6-PR-RT amplified products

3.1.2.1. Gel electrophoresis of single genome HIV-1 templates amplified by a nested NFL PCR

The wells from NFL PCR plates that fluoresced at the 15 copies per reaction target dilution in the plate viewing were separated by gel electrophoresis to determine whether a band of the correct size was present when exposed to ultraviolet light. Successful amplification with the modified NFL PCR approach using the newly designed primer in the nested PCR reaction from Katusiime *et al.* (2020) yields a band of interest at 9kb. Figure 3.2 shows an encircled band corresponding to a 9kb fragment in the lane labelled *5* and represents the results that were observed when NFL PCRs were repeated to produce more template. The wells containing the 9kb fragments were noted and the corresponding pre-nested products were used in the downstream ISLA assays and MDA amplifications.

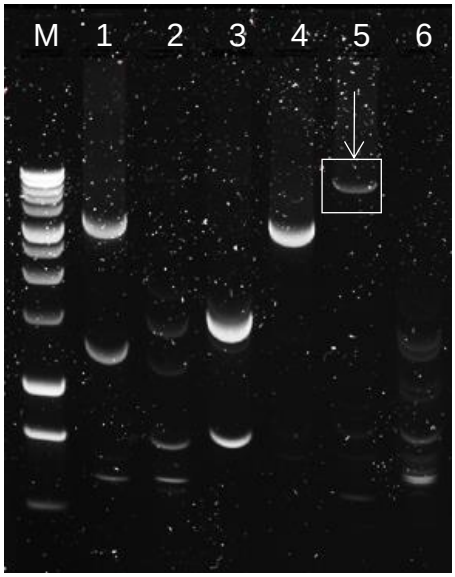

Figure 3.2: Gel photo showing an example of the results for a nested NFL PCR using 8E5 DNA as a template.

3.1.2.2. Gel electrophoresis of single genome HIV-1 templates amplified by a nested p6-PR-RT PCR

The wells from the nested p6-PR-RT PCR plates that fluoresced at the 1:27 target dilution when plate viewed were separated through gel electrophoresis to determine if successful amplification occurred. Successful amplification with the van Zyl *et al.* (2017) primer sets yields a fragment of 1.2kb. The results of the gel electrophoresis separation of the fluorescent wells present at the 1:27 target dilution are shown in Figure 3.3. All the lanes in Figure 3.3 contain a 1.2kb fragment. The wells containing the pre-nested products

corresponding to these nested amplicons were used as a template for the ISLA assays and MDA amplification.

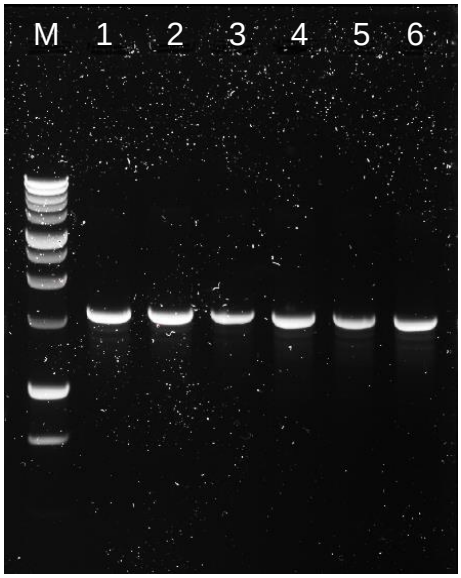

Figure 3.3: Gel electrophoresis results of the fluorescent wells from a nested p6-PR-RT amplification at 1:27 target dilution. *M*, 1kb molecular marker, *1 - 6*, nested p6-PR-RT amplicons diluted to 1:27 in 5 mM Tris-HCl.

3.2. Integration site loop amplification assay

3.2.1. 3'LTR ISLA assay

The ISLA assay was initially attempted using the pre-nested product of a provisionally intact HIV-1 patient sample that was amplified by the NFL PCR as a template. The template was replaced with extracted 8E5 DNA for optimisation purposes as patient samples are limited and provisionally intact samples are very rare. The 8E5 DNA was first diluted in HIV uninfected PBMC DNA as part of the dilution series and for these investigations, a negative control was included where DNA from HIV uninfected PBMCs replaced the diluted 8E5 DNA template. The 8E5 DNA was serially diluted in 5 mM Tris-HCl for subsequent optimisation investigations.

3.2.1.1. ISLA on a HIV-1 patient sample

A pre-nested provisionally intact HIV-1 patient sample initially served as a template for the 3'LTR ISLA assay. A non-template control and negative control were included in the amplification steps. The result of the initial attempt of the 3'LTR ISLA assay is shown in Figure 3.4. No unique bands were identified in the HIV-1 sample lane as all the bands corresponded to the bands that were observed in either the non-template control or the negative control lane.

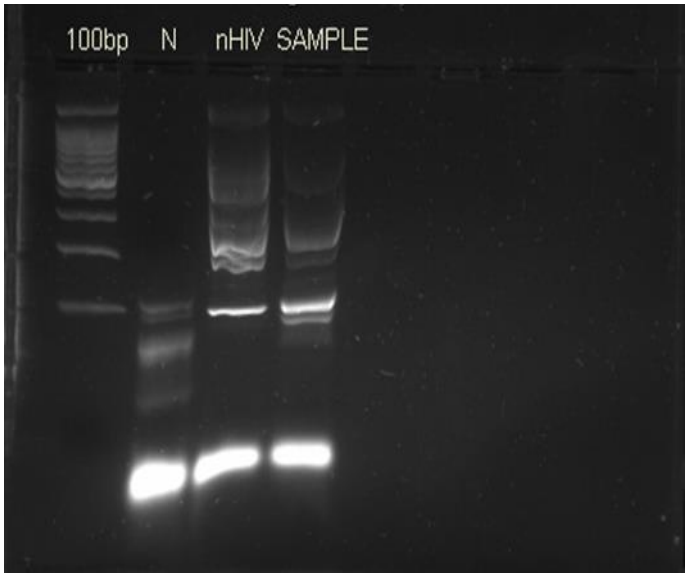

Figure 3.4: 3'LTR ISLA assay using the pre-nested product of a provisionally intact NFL HIV-1 sample and a negative control. *100 bp*, molecular marker, *N*, non-template control, *nHIV*, negative control (uninfected PBMCs), *SAMPLE*, provisionally intact NFL HIV-1 patient sample.

The products amplified in the two PCR steps that target the R region of HIV (*Step 4* and *Step 5* of ISLA) were individually observed through gel electrophoresis. The amplicons from both PCRs were viewed to determine if the second PCR targeted a specific product that was produced during the first round of PCR. A non-template control and a negative control were included and subject to amplification by the 3'LTR ISLA assay. Figure 3.5 shows the results of the two PCR amplifications for the 3'LTR ISLA assay. One unique band was present in the PCR 2 HIV-1 sample lane (encircled), however, due to the proximity to the band below, gel excision and purification of the band was not possible. In addition to this, the unique band occurred below the lowest molecular marker of 250/253 base pairs and the fragment was, therefore, shorter than the minimum 300 base pair limit set for Sanger sequencing of an ISLA product as a minimum of 30 base pairs of the HIV genome needs to be accurately sequenced prior to the junction between HIV and human genomic sequence to link integration sites in downstream analysis.

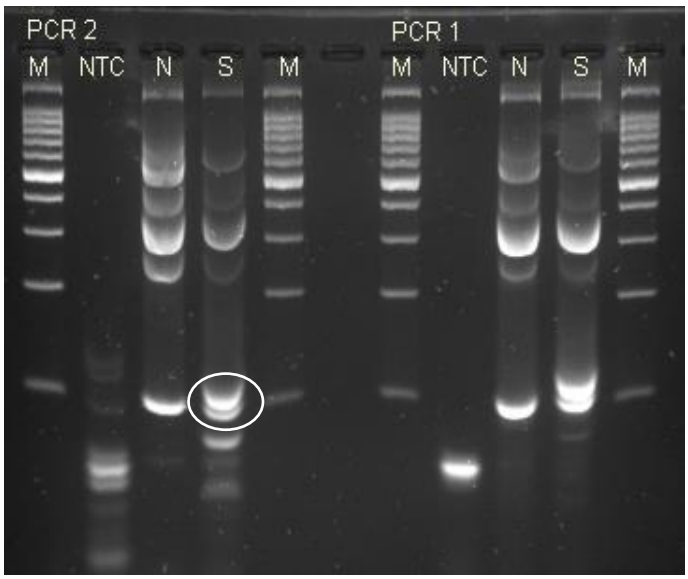

Figure 3.5: The first and second PCRs of the 3'LTR ISLA assay using a provisionally intact NFL HIV-1 patient sample. *M*, 1kb molecular marker, *NTC*, non-template control, *N*, negative control (uninfected PBMCs), *S*, ISLA on a provisionally intact HIV-1 patient sample.

3.2.1.2. ISLA on extracted 8E5 DNA

Extracted 8E5 DNA was diluted in the DNA of a HIV negative PMBC sample to the final HIV-1 DNA copies of 60, 30, 15, and 7,5 per reaction as part of a 2-fold dilution series to mimic a patient sample with a low concentration of HIV infected cells, amongst a background of cellular DNA. A negative control (uninfected PBMCs), a non-template control and 12 replicates of each final copy number were subject to NFL amplification and the plate viewing method was used to view the results. The plate view of the nested NFL PCR amplicons are shown in Figure 3.6. All the pre-nested products corresponding to the fluorescent wells in Figure 3.6 were used as templates for the 3'LTR ISLA assay. The effect of using various HIV copy numbers and therefore multiple HIV genomes as a template for the ISLA assay was investigated. The results of using 15 or 7,5 copies of HIV per reaction are shown in Figure 3.7. The 60 and 30 copies per reaction yield no unique fragments. Overall, for all the copy numbers, most of the fragments present corresponded to those present in the non-template or negative control lanes. A unique fragment of approximately 380 base pairs was observed in the first lane of the *7,5 cp/rxn* and is encircled in Figure 3.7. The unique fragment was gel excised and purified by the NucleoTraP®CR Kit for direct Sanger sequencing with the designed 3'LTR Junction-seq primer (section 3.2.9.).

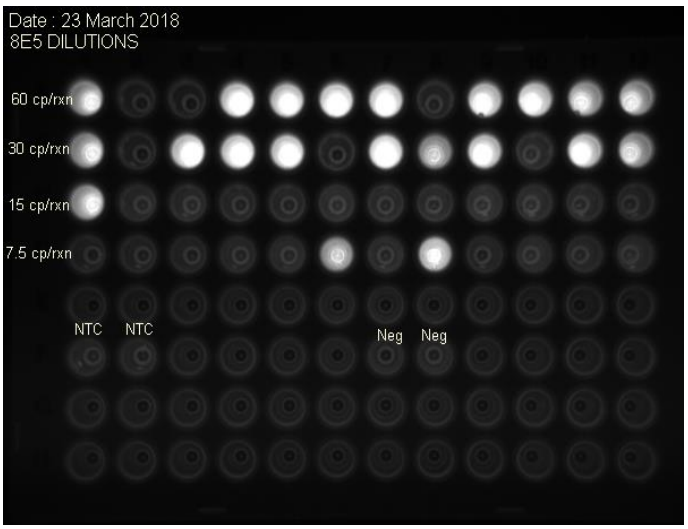

Figure 3.6: Plate viewing of nested NFL amplicons using 8E5 DNA that has been diluted to various copy numbers as a template. *60 cp/rxn*, 60 copies of HIV per reaction, *30 cp/rxn*, 30 copies of HIV per reaction, *15 cp/rxn*, 15 copies of HIV per reaction, *7,5 cp/rxn*, 7,5 copies of HIV per reaction, *NTC*, non-template control, *Neg*, negative control (uninfected PBMCs).

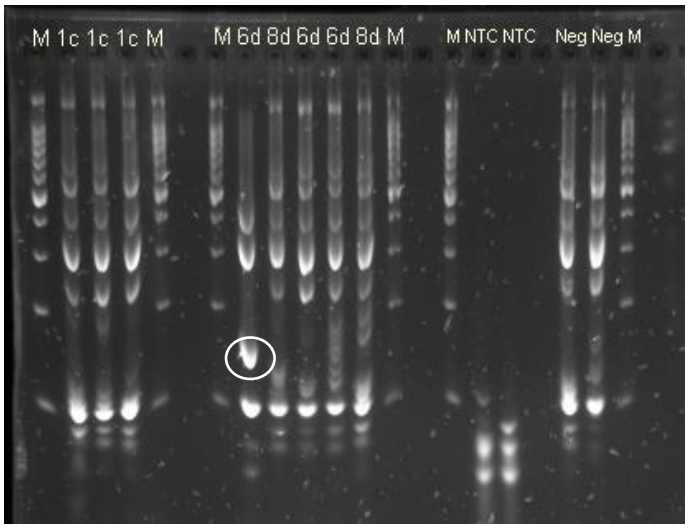

Figure 3.7: 3'LTR ISLA assay using NFL pre-nested 8E5 DNA amplified products as a template for ISLA. *1kb*, molecular marker, *NTC*, non-template control, *Neg*, negative control (uninfected PBMCs), *1c*, 8E5 DNA diluted to 15 copies per reaction, *6d/8d*, 8E5 DNA diluted to 7,5 copies per reaction.

3.2.2. 5'LTR ISLA assay

The 5'LTR ISLA assay approach was attempted after unsuccessful amplification was observed with the 3'LTR ISLA assay. The 5'LTR assay was initially investigated using the remaining 8E5 pre-nested templates generated by the NFL PCR shown in Figure 3.6. Figure 3.8 shows the results of using various copy numbers as a template for the 5'LTR ISLA assay. The encircled faint bands in Figure 3.8, which were brighter and clearer upon analysis, were present at the correct base pair length (350 – 400 base pairs) and were gel

excised and purified for sequencing using the NucleoTraP®CR Kit. The sequencing results are presented in section 3.2.9. The larger fragment encircled and present in the lane labelled B8 was the brightest of all the bands that were excised and resulted in the best purity ratio and was therefore cloned and sequenced (section 3.2.10.).

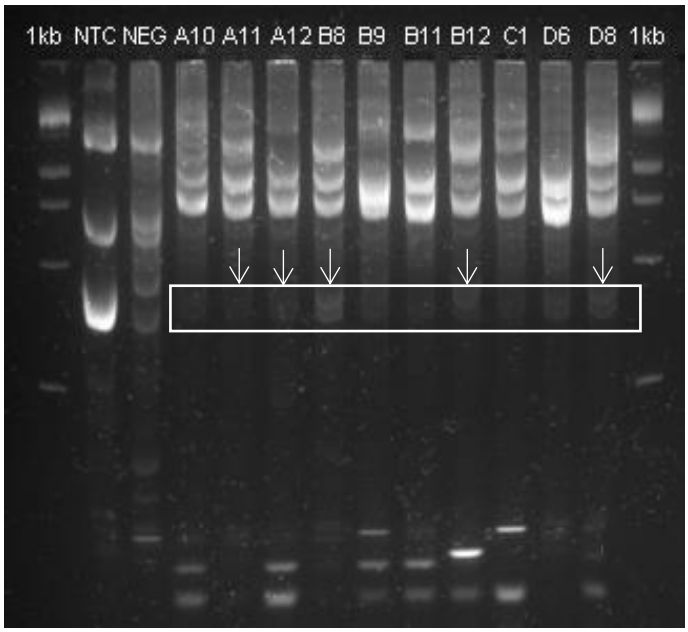

Figure 3.8: 5'LTR ISLA assay using pre-nested NFL amplified 8E5 DNA at various copy numbers per reaction as a template. *1kb*, molecular marker, *NTC*, non-template control, *Neg*, negative control (uninfected PBMCs), *A10 - A12*, 60 copies of HIV per reaction, *B8 - B12*, 30 copies of HIV per reaction, *C1*, 15 copies of HIV per reaction, *D6 - D8*, 7,5 copies of HIV per reaction.

3.2.3. Optimal annealing temperature

The Wagner *et al.* (2014) annealing temperatures were initially utilised in all ISLA assay approaches. Unsuccessful amplification across the 8E5 HIV-1 integration sites and the adjacent human genomic sequence, lead to the investigation of primer annealing temperatures. Two approaches to investigate the annealing temperature changes were established, the first ISLA approach used the Wagner *et al.* (2014) annealing temperatures and the second approach used lower annealing temperatures as described in section 2.3.3.2.4. The results of utilising the different annealing temperatures for *Step 4* and *Step 5* of ISLA are shown in Figure 3.9. The gel photo shows the Wagner *et al*. (2014) annealing temperatures on the left (*1 - 12*) and the lowered annealing temperatures on the right (*13 - 24*). Three unique and bright bands were observed in lanes *8*, *9* and *17* and are indicated by arrows in Figure 3.9 and measure at 170, 200 and 450 base pairs, respectively. The bands in the lanes labelled *8* and *9* were smaller than the 300-base pair limit set for ISLA products to be Sanger sequenced but were the longest fragments observed in the first

approach and were investigated further. All three bands were gel excised and purified using the NucleoTraP®CR Kit for Sanger sequencing. In addition to this, the purified bands were cloned into the pJET1.2/blunt plasmid vector and sequenced.

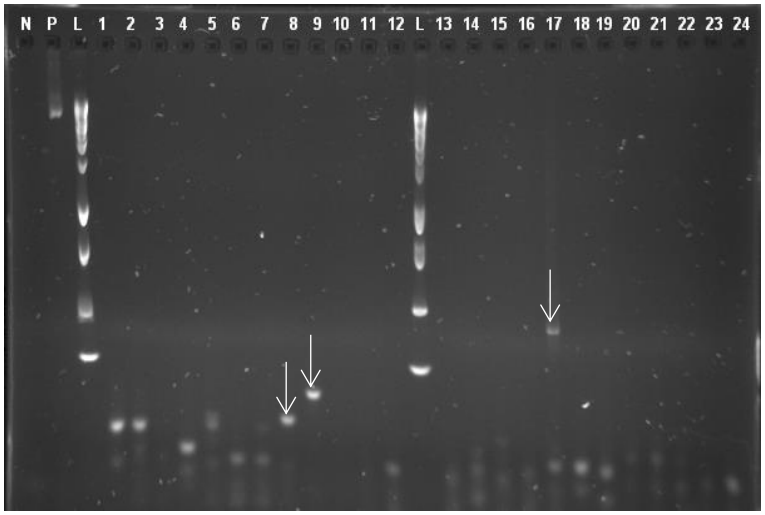

Figure 3.9: 5'LTR ISLA assay using p6-PR-RT pre-nested products as a starting template. *N*, negative control, *P*, positive control, *L*, 1kb molecular marker, *1 - 12*, Wagner *et al.* (2014) annealing temperatures, *13 - 24*, lowered annealing (modified) temperatures.

3.2.4. Increasing the ISLA template concentration

The concentration of the starting template for ISLA was increased by MDA. The effect of using linear products amplified by MDA as a template for the ISLA assay was investigated. Two investigations were conducted to determine the effect of using MDA.

The first investigation used pre-nested NFL products at target dilution, 15 copies per reaction, followed by MDA amplification. The 8E5 DNA was diluted to target in 5 mM Tris-HCl for this experiment. The results of this attempt are shown in Figure 3.10. Three unique, very bright and large sized fragments measuring approximately 2 500 base pairs in length were observed and are encircled in the figure, although the fragments were much larger than the expected fragment length for ISLA, approximately 300-base pairs, they were selected for further investigation to determine their significance. No bands were present in the negative control lane which further encouraged the decision to sequence the bands present in lanes, *5, 8a* and *19*. The fragments were gel excised and purified with the NucleoTraP®CR Kit for direct Sanger sequencing. Samples *8a* and *19* were cloned into the CloneJET™ pJET1.2/blunt plasmid vector to determine their significance. The sequencing results are discussed in sections 3.2.9. and 3.2.10.2.2.

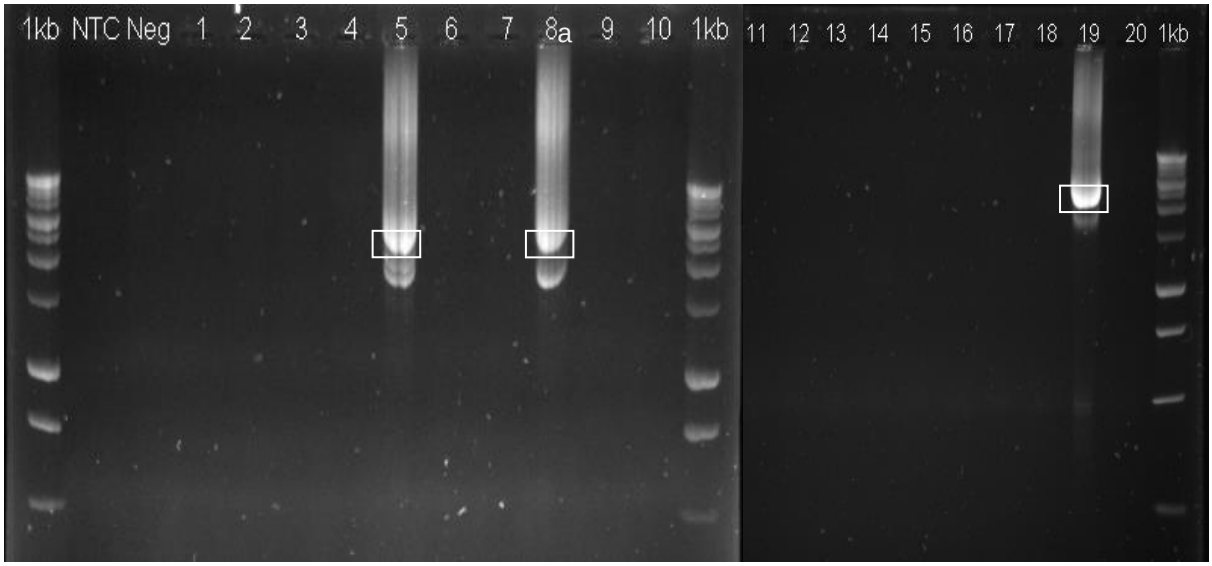

Figure 3.10: 5'LTR ISLA assay using 8E5 DNA at 15 copies per reaction target dilution. *1kb*, molecular marker, *NTC*, non-template control, *Neg*, negative control (uninfected PBMCs), *1 - 20*, 8E5 DNA diluted in Tris-HCl to 15 copies per reaction.

The second investigation utilised two approaches using the same initial 8E5 NFL and/or p6-PR-RT pre-nested products. The pre-nested products from the NFL and p6-PR-RT PCRs at their respective target dilutions were divided into two separate tubes to investigate the approaches. The first approach used the pre-nested products that were amplified by MDA (section 2.3.3.2.5.) prior to *Step 2* also known as the random priming step of ISLA. The second approach that was investigated, used the original pre-nested products as templates for ISLA. The results of the two approaches were compared and are shown in Figure 3.11. Both the approaches using the pre-nested NFL or p6-PR-RT products at their respective target dilutions as a template showed no amplification.

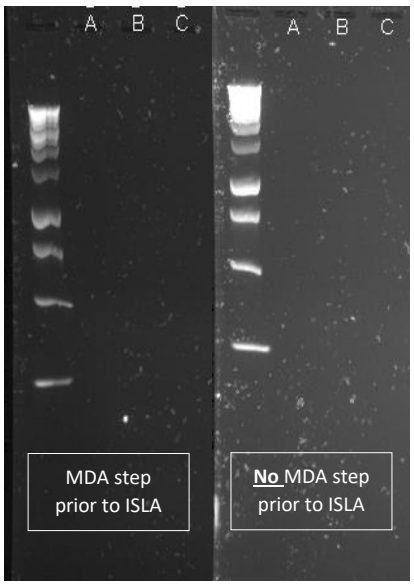

Figure 3.11: The result of using MDA amplified pre-nested products as a template versus using the pre-nested products directly as a template for ISLA. *A - C* (left-hand side), pre-nested NFL templates were amplified with MDA prior to ISLA, *A - C* (right-hand side), NFL pre-nested products were directly used as templates for ISLA.

3.2.5. Random primer concentration optimisation

Different random primer concentrations, 2 µM, 1.5 µM, 1 µM and 0.75 µM, were investigated for *Step 2* of ISLA. The 5'LTR ISLA assay was attempted using 8E5 pre-nested NFL and p6-PR-RT products at their respective target dilutions as templates. In addition to this, the influence of the MDA amplification prior to *Step 2* of ISLA was also investigated. Therefore, two different approaches for investigating the impact of using different random primer concentrations and MDA were established. The first approach used 8E5 pre-nested NFL or p6-PR-RT products at target dilution as templates for ISLA that were randomly primed using different primer concentrations. The second approach used 8E5 pre-nested NFL or p6-PR-RT templates that were amplified by MDA prior to the random priming step of ISLA. The use of different random primer concentrations was investigated in triplicate for both approaches. The results of the two approaches are shown in Figures 3.12 and 3.13. No amplification was observed in either approach.

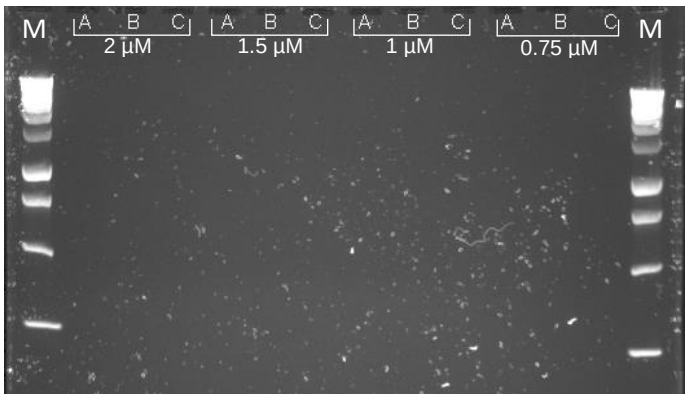

Figure 3.12: The first approach investigating different random primer concentrations and no MDA amplification of the templates. *M*, 1kb molecular marker, *2 µM, 1.5 µM, 1 µM, 0.75 µM,* random primer concentrations investigated in triplicate (*A - C*).

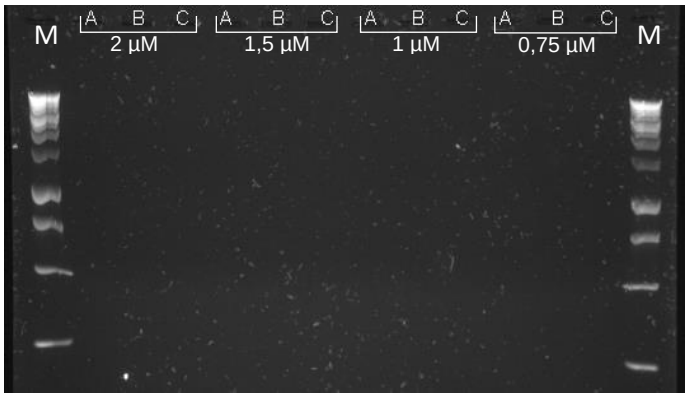

Figure 3.13: The second approach investigating the use of different random primer concentrations on MDA amplified templates. *M*, 1kb molecular marker, *2 µM, 1.5 µM, 1 µM, 0.75 µM,* random primer concentrations investigated in triplicate (*A - C*).

3.2.6. Comparison of DNA polymerases for ISLA

MyTaq™ Mix replaced RANGER Mix in the ISLA assay reactions to compare the efficacies of the two enzymes. The addition or exclusion of the MDA step prior to *Step 2* of ISLA was also investigated. Therefore, several 8E5 pre-nested p6-PR-RT products at target dilution were subject to MDA amplification and used as a template to investigate ISLA with the MyTaq™ enzyme mix. An additional amplification was simultaneously conducted using the pre-nested products as a template with no MDA amplification. The comparison of the two approaches is shown in Figure 3.14.

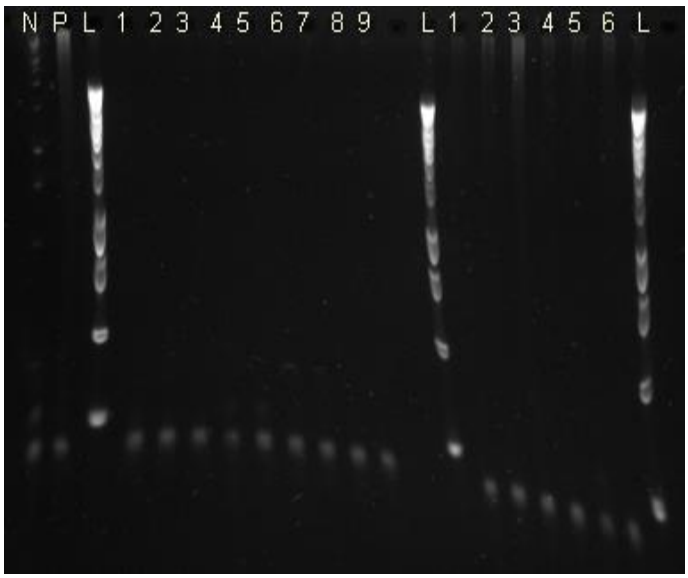

Figure 3.14: Results of ISLA with the MyTaq™ Mix in place of RANGER Mix. *N*, negative control, *P*, positive control which was 100 ng/µl of 8E5 DNA, *L*, 1kb molecular marker, *1 - 9*, MDA amplified ISLA template, *1 - 6*, No MDA prior to ISLA assay.

3.2.7. Subtype B ISLA assay

The subtype B ISLA assay published by Wagner *et al*. (2014) was attempted. Extracted 8E5 DNA was linearly amplified as described in section 2.3.3.2.8.and the linear products served as a template for the published ISLA method. ISLA was attempted on 24 linearly amplified products. The results of the subtype B ISLA assay are shown in Figure 3.15. No successful amplification was observed.

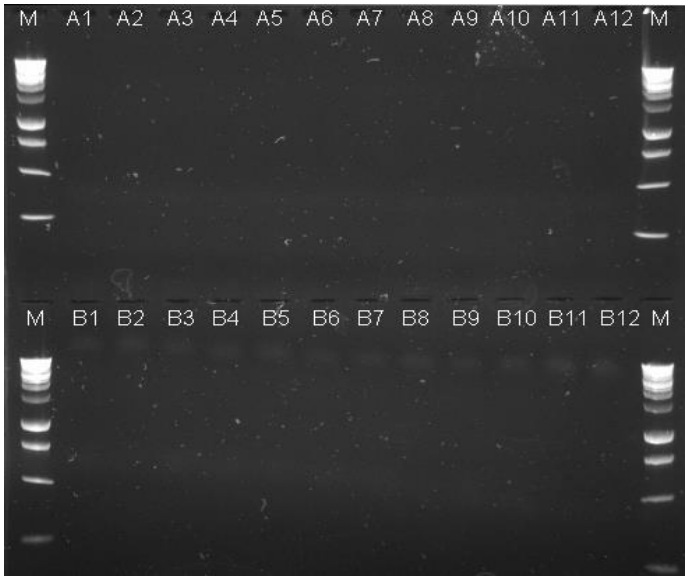

Figure 3.15: Results of the published Wagner *et al*. (2014) ISLA assay using linearly amplified 8E5 DNA as a template. M, 1kb molecular marker, *A1 - B12*, linearly amplified 8E5 DNA template.

3.2.8. Purification of ISLA products for sequencing

The unique ISLA products selected for sequencing were highlighted in the previous sections. All of the products were gel excised and purified using the NucleoTraP®CR Kit. The absorbance of the purified products were measured with the NanoDrop™ 1000 Spectrophotometer to determine their respective concentrations and purities and are shown in Table 3.1. As nucleic acids absorb at a wavelength of 260nm and proteins at a wavelength of 280nm, a ratio of absorbance at 260nm and 280nm (260/280) was used to assess the purity of the nucleic acid and a ratio measuring approximately 1.8 corresponds to 'pure' DNA and a ratio measuring approximately 2.0 is considered to be 'pure' for RNA.

3.2.9. Direct sequencing of ISLA products by Sanger sequencing

All the purified ISLA products presented in Table 3.1 were directly sequenced by Sanger sequencing with the designed 3'LTR Junction_seq or 5'LTR Junction_seq primers at CAF. The sequencing primer chosen was dependant on the ISLA approach that was utilised and are indicated in Table 3.1. A pGEM® positive control was included and successfully sequenced in all reactions, indicating that the sequencing reaction set up and sequencing conditions were optimal. The lengths of the resulting sequences ranged from 46 – 486 base pairs with an average length of 243 base pairs. ISLA samples 6d, B8, B12, 8 and 19 were excluded from further analyses as the designed 3'LTR and 5'LTR sequencing primers did not sequence these samples. In addition to the aforementioned samples, ISLA samples 8 and 9 from Figure 3.10, were also excluded as both sequences had <100 base pairs. Therefore, all the remaining samples >100 base pairs were used to perform NCBI's blastn followed by an alignment to the HIV Consensus C and HXB2 reference genomes, respectively. ISLA sample 5 was the only sample that returned a blastn result and showed an 87.13% percentage identity to a *Sporobolomyces* species available on GenBank (accession number: MG470907.1). No contigs were formed through sequence alignment to HIV Consensus C or HXB2.

Table 3.1: NucleoTraP®CR gel excision purification results and the primer used for sequencing for each ISLA product

ISLA approach	**Sample name**	**Figure**	DNA concentration (ng/µl)	260/280	260/230	**Sequencing primer name**
3'LTR	6d	3.7	50.1	1,92	1,14	3'LTR Junction_seq
5'LTR	A11 (lower band)	3.8	39,8	1,87	1,11	5'LTR Junction_seq
	A12 (lower band)		38,1	1,86	1,23	
	B8 (lower band)		68,0	1,83	1,28	
	B8		50,0	1,89	1,12	
	B12		75,4	1,87	1,23	
	D8		67,3	1,76	1,03	
	5	3.10	44,6	2,06	1,41	
	8a		32,9	2,35	1,09	
	19		82,6	1,94	1,21	
	8*	3.9	17,9	1,35	0,70	
	9*		14,1	1,41	0,64	
	17*		12,1	4,38	0,68	

* average of three readings

Cloned samples

3.2.10. Cloning of ISLA assay products

The primer binding sites of the 3'LTR Junction_seq and the 5'LTR Junction_seq sequencing primers occurred near the integration site thereby making recovery and identification of these sites unattainable. Several purified ISLA products were selected to be cloned into a plasmid vector and sequenced by Sanger sequencing. The purified ISLA products were ligated into a kit plasmid vector following the InsTAclone™ or the CloneJET™ cloning method (sections 2.3.3.6.2. and 2.3.3.6.3.) The names and sizes of the purified ISLA products chosen for cloning are shown in Table 3.2 together with their respective cloning kit and sequencing primers.

Table 3.2: ISLA samples that were cloned and sequenced

Figure number	**Sample name** (according to figure)	**Insert size in base pairs** (approximate)	**Cloning kit**	**Sequencing primers**
3.8	B8	450	InsTAclone™	M13/pUC Forward and M13/pUC Reverse
3.10	8a	2 500	CloneJET™	pJET1.2 Forward Sequencing Primer and pJET1.2 Reverse Sequencing Primer
	19	2 500		
3.9	8	170		
	9	200		
	17	450		

3.2.10.1. Successful ligation confirmation of the fragment of interest into the kit plasmid vector

3.2.10.1.1. InsTAclone™ PCR Cloning Kit

ISLA product B8 was ligated into the pTZ57R/TA vector as described in section 2.3.3.6.2.1. The transformed bacterial cells were divided into three volumes and spread across the surface of a prepared LB agar plate as described in section 2.3.3.6.2.2. The 3 plates were observed for the presence of single colonies following the overnight incubation at 37°C. A single white-creamy colony was selected and grown in a bacterial culture overnight and

purified with the GeneJET Plasmid Miniprep kit (2.3.3.6.2.5.). The purified plasmid was diluted with nuclease-free water to 20 ng/µl for the DNA sequencing reaction and was sequenced by Sanger sequencing.

3.2.10.1.2. CloneJET™ PCR Cloning Kit

A total of 5 ISLA products, namely sample 8a, 19, 8, 9 and 17, were selected to be ligated into the pJET1.2/blunt cloning vector of the CloneJET™ PCR Cloning Kit. The method described in section 2.3.3.6.3.1. was followed to ligate the ISLA products into the cloning vector. Two selected volumes of the transformed bacteria (2.3.3.6.3.2.) for each ISLA product was spread across the surface of two LB agar plates. The plates were viewed following the overnight incubation at 37°C for the presence of single colonies. A colony PCR with a high-fidelity enzyme was performed on the selected colonies to confirm the presence of the ISLA products. The results of the colony PCR are shown in Figures 3.16, 3.17 and 3.18. The colony PCR reactions from the fragments encircled in the figures were diluted in nuclease-free water to 20 ng/µl and sequenced by Sanger sequencing.

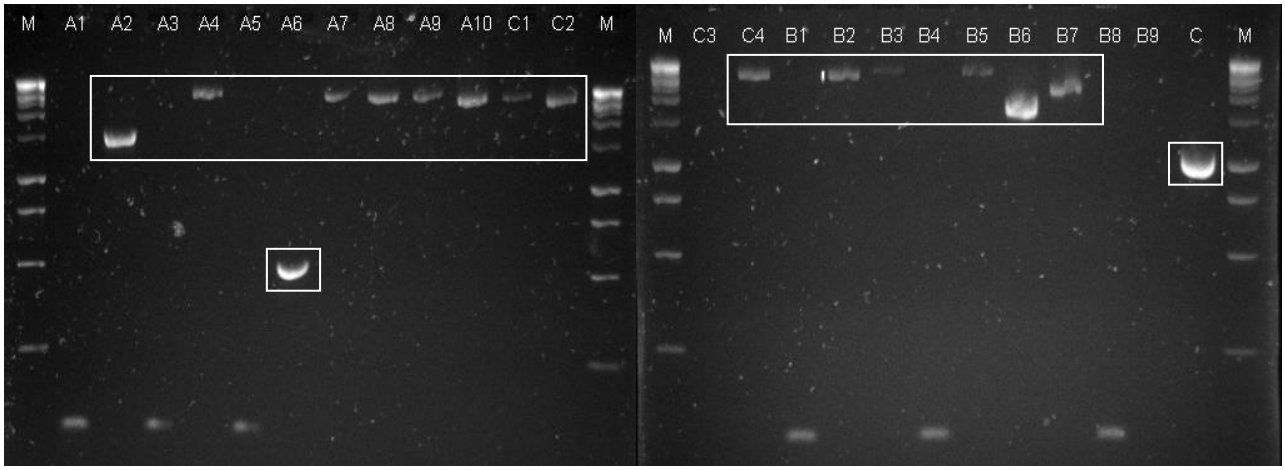

Figure 3.16: Colony PCR results of ISLA samples 8a and 19. *M*, 1kb molecular marker, *A1 - A10* and *C1 - C4*, single colonies selected from the plate inoculated ISLA sample 8a, *B1 - B9*, single colonies selected from the plate inoculated ISLA sample 19, *C*, CloneJET™ PCR Cloning Kit positive control.

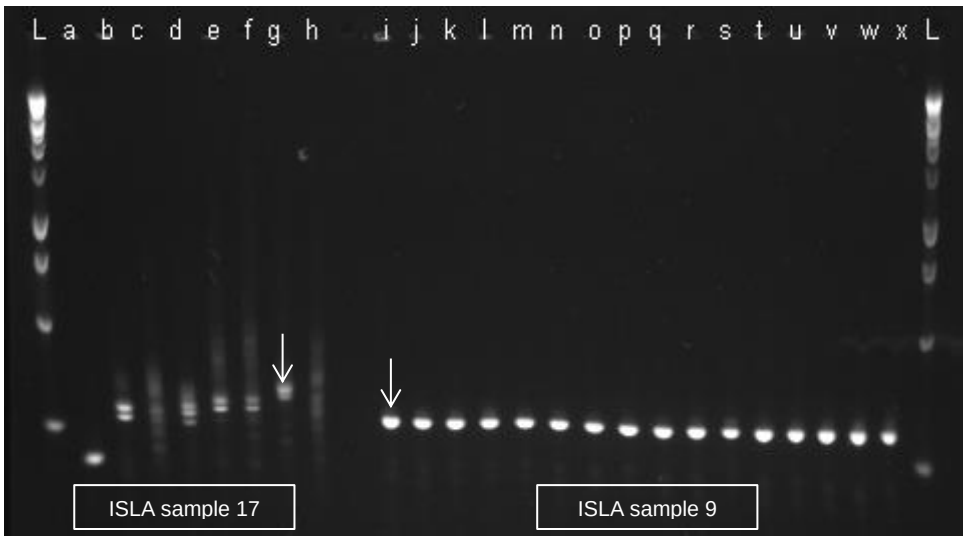

Figure 3.17: Colony PCR of ISLA samples 17 and 9 following the CloneJET™ PCR cloning procedure. *L*, 1kb molecular marker, *a* - *h*, ISLA sample 17 single colonies, *i* - *x*, ISLA sample 9 single colonies.

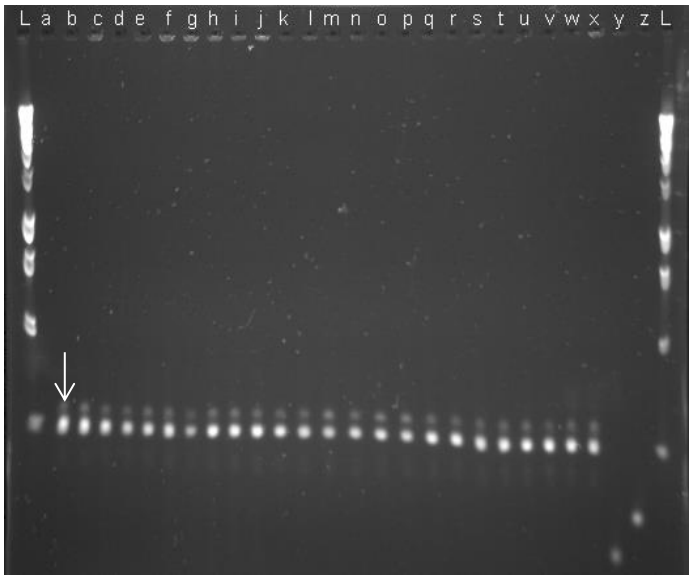

Figure 3.18: Colony PCR results for ISLA sample 8 cloned by the CloneJET™ PCR Cloning Kit. *L*, 1kb molecular marker, *a* - *z*, single colonies selected from both inoculated LB agar plates.

3.2.10.2. Sanger sequencing results of ISLA products cloned by the InsTAclone™ or CloneJET™ PCR cloning kits

3.2.10.2.1. InsTAclone™

The forward and reverse sequencing of ISLA sample B8 was conducted in triplicate. A MUSCLE alignment with 100 iterations was done for the 3 forward and 3 reverse sequences, respectively, to generate a raw consensus sequence for each direction. The generated raw consensus sequences were aligned to each other resulting in an overlap of 350 base pairs shown in Figure 3.19. The forward and reverse raw consensi were individually loaded onto

NCBI's blastn and both matched to the Eukaryotic synthetic construct chromosome 16 nucleotide sequence (accession number: CP034494.1, percentage identity: 99.19% for forward and 98.63% for reverse consensus sequences). In addition, the consensi were aligned to the HIV Consensus C and HXB2 reference genomes and no contigs were identified.

Figure 3.19: MUSCLE alignment of the raw forward and reverse consensus sequences of ISLA sample B8. Grey overlapping sections indicate agreements between the consensi and the presence of colour within the grey sections indicate disagreements between the two consensi.

3.2.10.2.2. CloneJET™

The 15 colony PCR fragments highlighted in Figure 3.16 from ISLA samples 8a and 19, were sequenced the pJET sequencing primer set in a forward and reverse sequencing reaction. Sequences with <100 base pairs were excluded from further analyses. NCBI's blastn was performed on each raw sequence, followed by an alignment to HIV Consensus C. All the blastn results except for sample 8a C1 forward, returned matches to various HIV-1 sequences available on GenBank. A total of 23 sequences matched to HIV-1 sequences, nine matched to accession number MN691959.1 (percentage identity average: 93.31%), seven matched to MN467309.1 (93.5%), three matched to MN989412.1 (97.31%) and the remaining four sequences returned matches to MN685352.1(96.30%), MN090293.1 (82.86%), KX129262.1 (87.22%) and Z11530.1 (90.03%), respectively. A MUSCLE alignment with 100 iterations was performed with all the raw forward sequences. Cloned samples A4, A7, A8, A9, B5, B7 and C2 showed good agreement in overlapping regions and a forward consensus of 976 base pairs with 6 gaps was generated and aligned to the HIV Consensus C and HXB2 reference sequences. The forward consensus sequence aligned to HXB2 at 8604 - 9580 base pairs. The reverse raw consensus sequence was generated in the same way as the forward consensus sequence, however, cloned samples A4, B6, C1, C2 and C4 showed the most similarity across overlapping regions and a 946 base pair consensus sequence with 6 gaps was generated. The raw reverse consensus sequence aligned to HXB2 at 6353 - 7292 base pairs with 6 gaps. The results of the alignments to the HXB2 reference genome are shown in Figure 3.20. and the raw forward and reverse consensi do not overlap. All the sequences contained the NFL HIV-1 PCR target

sequences but no human genome, therefore no integration sites were identified. Many gaps, insertions and deletions, and disagreements between the consensi and the reference genome were observed in the alignments to HIV Consensus C.

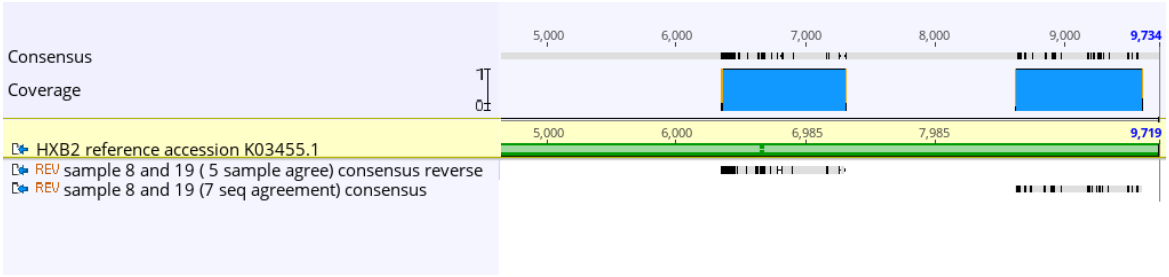

Figure 3.20: Alignment of the raw forward and reverse consensus sequences to the HXB2 reference genome. HXB2 reference genome is indicated in green and the areas where the two consensi best align to the reference genome are indicated in blue.

Three resulting fragments from the colony PCR of cloned ISLA samples 8, 9 and 17 were selected for sequencing and are highlighted in Figures 3.17 and 3.18. The resulting raw sequences obtained by Sanger sequencing were subject to NCBI's blastn and ISLA samples 9 and 17 a returned a "no significant similarity" result. However, for ISLA sample 8, the blastn of the raw forward and reverse sequences showed a 100% percentage identity to a cloning vector sequence on GenBank (accession number: KP223705.1). The raw sequences were also aligned to HIV Consensus C and HXB2 and no contigs were identified.

3.3. ONT sequencing results

3.3.1. Premium whole genome amplification

3.3.1.1. Purification of MDA amplified products

Multiple MDA reactions were conducted using a 8E5 NFL pre-nested product as a template (referred to as the NFL MDA amplified product in the following sections) for optimisation. Several wells containing the NFL MDA amplified product were combined into a 1.5 ml tube and was used to investigate all methods of purification. This allowed for the comparison between methods and the determination of the best method of purification. Prior to purification, the concentration and purity of the combined NFL MDA amplified product was measured using the NanoDrop™ 1000 Spectrophotometer and the Qubit® 2.0 Fluorometer and the result is shown in Table 3.3 below.

Table 3.3: Concentration and purity of the 8E5 NFL MDA amplified product prior to purification

	NanoDrop™ 1000 spectrophotometer			Qubit® 2.0 Fluorometer	
	DNA concentration (ng/µl)	260/280▪	260/230	DNA concentration (ng/ml)	Total concentration (ng)*
NFL MDA product	623,6	1,90	1,72	2320	113,68

▪ Ratio of ~1.8 corresponds to 'pure' DNA and a ratio of ~2.0 corresponds to 'pure' RNA

Several methods for purifying the NFL MDA amplified product were investigated and have been described in section 2.3.4.3.1. Table 3.4 shows the results of these purification methods. Due to the concentration and volume requirements for downstream analyses, a limited amount of purified sample remained for the quality and quantity determination. Therefore, where sufficient product remained both NanoDrop™ 1000 Spectrophotometer and Qubit® 2.0 Fluorometer measurements were taken and for purified samples with limited volume remaining only a Qubit® 2.0 Fluorometer measurement was obtained.

Table 3.4: Methods used to purify MDA amplified products and the respective results

		NanoDrop™ 1000 spectrophotometer			Qubit® 2.0 Fluorometer	
Purification method		DNA concentration (ng/µl)	260/280▪	260/230	DNA concentration (ng/ml)	Total concentration (ng)*
AMPure XP bead purification (2.3.4.3.1.1)					94	4,61
					741	36,31
Ethanol precipitation	Qiagen (2.3.4.3.1.2.1.)	335,5	1,88	2,38	1122	54,98
	MRC Holland (2.3.4.3.1.2.2.)	339,7	1,82	2,24	2060	100,94
	In-house (2.3.4.3.1.2.3.)	140,5	1,68	2,28	2020	98,98

*total ng in the remaining 49µl for downstream analysis

▪ Ratio of ~1.8 corresponds to 'pure' DNA and a ratio of ~2.0 corresponds to 'pure' RNA

3.3.1.2. T7 Endonuclease I digestion

3.3.1.2.1. Optimisation of T7 Endonuclease I digestion

The MRC Holland ethanol precipitation method was identified as the best method for purifying NFL MDA amplified products as shown in Table 3.4. This purification method was therefore used to purify sufficient NFL MDA amplified product to optimise the T7 Endonuclease I digestion step in ONT's Premium whole genome amplification protocol. As mentioned in section 2.3.4.3.2, an initial input concentration of 1.5 µg of the purified NFL MDA amplified product was investigated, however, a rapid decrease in the concentration of the product following the digestion with the T7 Endonuclease I enzyme was measured and the alternative method using an input concentration of 3.5 µg was investigated with an adjusted incubation step based on the recommendations of other ONT users. The resulting concentrations of T7 Endonuclease I digested products were determined by the NanoDrop™ 1000 Spectrophotometer and are shown in Table 3.5. Overall, a significant loss in concentration was observed with both input concentrations.

Table 3.5: Concentration of T7 Endonuclease I digested products

Input concentration of purified NFL MDA amplified product	DNA concentration (ng/µl)	260/280▪	260/230
1.5 µg	86,0	1,44	0,92
3.5 µg	127,2	1,55	1,15

▪ Ratio of ~1.8 corresponds to 'pure' DNA and a ratio of ~2.0 corresponds to 'pure' RNA

3.3.1.2.2. Purification of T7 Endonuclease I digested products

The products from both 1.5 µg and 3.5 µg input concentrations used in the T7 Endonuclease I digestion attempts were used to investigate various purification methods. The methods utilised in the purification of these digested products were described in section 2.3.4.3.3. Initial purification attempts were investigated using the digested product from the 1.5 µg input concentration reaction, however, due to low recovery following purification, the method was changed to the 3.5 µg input concentration reaction. Therefore, ONT's recommended purification method using AMPure XP beads with a custom buffer system (section 2.3.4.3.3.1) was investigated using both T7 Endonuclease I digested products as it yielded the best results for the initial 1.5 µg input concentration reaction. Table 3.6 shows the various

purification methods that were used with the respective input concentrations with the resulting concentrations and purities of the products for each method measured by the NanoDrop™ 1000 Spectrophotometer. Furthermore, since the MRC Holland ethanol precipitation method gave the best results in the NFL MDA amplified product purification, the method was investigated with an overnight incubation at -20°C.

Table 3.6: Purification method results used on T7 Endonuclease I digested products

1.5 µg purified DNA input			
Purification method	DNA concentration (ng/µl)	260/280▪	260/230
AMPure XP with custom buffer (2.3.4.3.3.1.)	0,6	0,87	Invalid
	1,3	1,01	1,54
	1,2	2,56	0,84
MRC Holland Ethanol precipitation (2.3.4.3.1.2.2.)	0,2	Invalid	0,18
NucleoSpin® (2.3.4.3.3.5.)	3,3	2,18	0,27
3.5 µg purified DNA input			
Purification method	DNA concentration (ng/µl)	260/280▪	260/230
AMPure XP with custom buffer (2.3.4.3.3.1.)	0,3	Invalid	Invalid
AMPure XP bead purification (2.3.4.3.3.2.)	Invalid	0,86	0,79
MRC Holland Ethanol precipitation (-20°C incubation) (2.3.4.3.3.3.)	11,6	2,36	1,90
NucleoTraP®CR (2.3.4.3.3.4.)	0,7	0,64	0,37

Invalid indicate blank readings that were higher than sample readings possibly due to the presence of ethanol
▪ Ratio of ~1.8 corresponds to 'pure' DNA and a ratio of ~2.0 corresponds to 'pure' RNA

3.3.1.3. ONT sequencing of the NFL MDA amplified product

Additional 8E5 NFL MDA amplified product was generated, the 3.5 µg T7 Endonuclease I digestion method was used to remove hyperbranched structures and the products were purified using the best methods of purification prior to DNA preparation and sequencing on the GridION. The MRC Holland ethanol precipitation method was used to purify the NFL

MDA amplified product and the T7 Endonuclease I digested product, for the latter, the adapted method using the overnight incubation at -20°C was used instead of the 15-minute incubation on ice. The concentrations of the resulting products were measured with the NanoDrop™ 1000 Spectrophotometer after various steps to determine where significant losses in concentration and purity occurred and to determine whether further processing of the product could occur for ONT sequencing. The resulting concentrations and purities after the respective purification steps are displayed in Table 3.7.

Table 3.7: Concentrations and purity of the product after various steps of purification

Sample name	DNA concentration (ng/µl)	260/280▪	260/230
NFL MDA product	542.3	1.88	1.56
Before T7 Endonuclease I digestion			
Purified NFL MDA products	737	1.77	2.20
	786.6	1.79	2.25
After T7 Endonuclease I digestion			
Purified T7 Endonuclease I digested products	5.5	1.74	5.37
	2.4	1.15	Invalid

Invalid indicate blank readings that were higher than sample readings possibly due to the presence of ethanol
▪ Ratio of ~1.8 corresponds to 'pure' DNA and a ratio of ~2.0 corresponds to 'pure' RNA

Further investigation of the NFL MDA amplified product was done to determine if the fragment size corresponded to the expected 9kb fragment length by separation through gel electrophoresis. The result of the gel electrophoresis separation is shown in Figure 3.21. No band was visible upon analysis and the gel electrophoresis step was repeated for a second time under the same conditions and a third time using an E-Gel™ Agarose Gels with SYBR™ Safe DNA Gel Stain, 1.2% (Invitrogen, USA) to confirm the result. All three investigations yielded the same result. Due to the insufficient concentration of the purified T7 Endonuclease I digested product, in conjunction with no visible 9kb fragment, ONT sequencing of the NFL MDA product did not proceed.

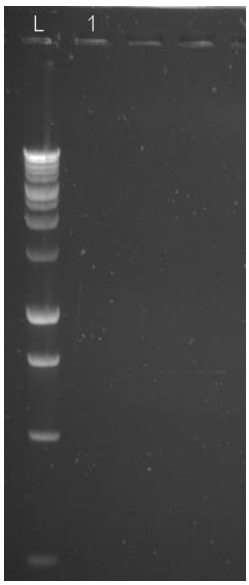

Figure 3.21: The NFL MDA amplified product was not visible following separation by gel electrophoresis and exposure to UV light. *L*, 1kb molecular maker, *1*, NFL MDA amplified 8E5 product.

ONT's Premium whole genome amplification protocol appeared to be the ideal method for identifying the integration sites of HIV-1. Although the processing of the 8E5 NFL MDA amplified product yielded unexplained results, the integration site of a unique patient sample needed to be identified for a PhD student project. This sample was first amplified by the NFL PCR to identify single genome amplified products using Poisson's distribution. The pre-nested NFL products corresponding to wells containing single genome amplified products were MDA amplified and processed following the Premium whole genome amplification protocol to be sequenced with the GridION. The optimised methods and purification steps were attempted on the patient sample and were compared to the 8E5 NFL MDA amplified product at various steps. The MDA amplification of the patient sample was conducted in quadruplicate and are labelled as *NFL 1*, *NFL 2*, *NFL 3* and *NFL 4*. The 8E5 and patient MDA amplified products were visualised by using the plate viewing method described in section 2.3.2.3.3.1. and the concentrations and purities of the products were measured with the NanoDrop™ 1000 Spectrophotometer prior to T7 Endonuclease I digestion. Due to the losses in concentration previously observed from the purification of the 8E5 NFL MDA amplified product to the purification of T7 Endonuclease I digested products, the initial purification of the patient MDA amplified products was omitted in an attempt to limit losses in concentration. The pre-nested p6-PR-RT MDA amplified product that will be described in section 3.3.1.4. was also included in the comparison. The concentration and purity results measured by the NanoDrop™ 1000 Spectrophotometer of the various samples are shown in Table 3.8 and Figure 3.22. Patient samples *NFL 2* and *NFL 4* had the best

recoveries and the MDA amplified product and T7 Endonuclease I digested products for these samples were separated by gel electrophoresis to determine the size of the product and the results are shown in Figure 3.23.

Table 3.8: Comparison of 8E5 NFL MDA amplified product and patient NFL MDA amplified sample pre-purification and post T7 Endonuclease I digestion purification

	MDA amplified samples			**Post T7 Endonuclease I digestion**		
8E5 samples	DNA concentration (ng/µl)	260/280▪	260/230	DNA concentration (ng/µl)	260/280▪	260/230
NFL	528,4	1,89	1,57	-	-	-
p6-PR-RT	621,5	1,90	1,71	Results in Table 3.9 section 3.3.1.4		
Patient sample identifiers						
NFL 1	861,7	1,47	1,39	8,2	1,90	5,31
NFL 2	491,6	1,84	1,57	14,6	1,77	2,24
NFL 3	487,0	1,84	1,57	14,0	2,19	4,04
NFL 4	483,1	1,84	1,55	14,8	1,87	2,86

▪ Ratio of ~1.8 corresponds to ‘pure’ DNA and a ratio of ~2.0 corresponds to ‘pure’ RNA

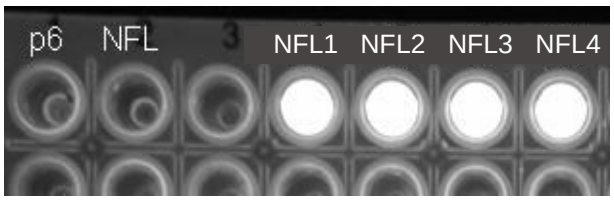

Figure 3.22: Plate view of the 8E5 NFL and p6-PR-RT MDA amplified products and the patient NFL MDA amplified products. *p6*, 8E5 p6-PR-RT MDA amplified product, *NFL*, 8E5 NFL MDA amplified product, *NFL1 - NFL4*, patient NFL MDA amplified products.

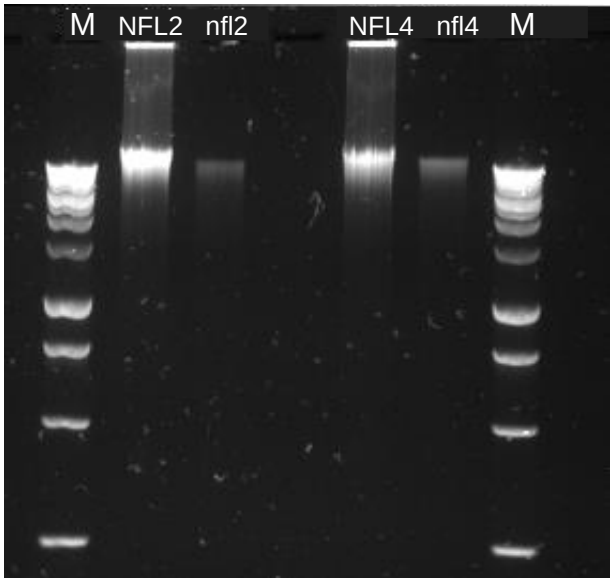

Figure 3.23: Patient NFL MDA amplified sample and T7 Endonuclease I digested sample. *M*, 1kb molecular marker, *NFL2/NFL4*, pre T7 Endonuclease I digestion, *nfl2/nfl4*, post T7 Endonuclease I digestion.

3.3.1.4. Sequencing preparation of the p6-PR-RT MDA amplified product

Extracted 8E5 DNA was amplified by the p6-PR-RT PCR and the pre-nested product corresponding to the nested wells that fluoresced under ultraviolet light were amplified by MDA (referred to as p6 MDA amplified product in the sections below) in an attempt to successfully recover the integration site of 8E5. The plate viewing of the p6 MDA amplified product is shown in Figure 3.22. ONT's Premium whole genome amplification protocol was utilised and the p6 MDA amplified product replaced the NFL MDA amplified product as a template in the Premium whole genome amplification protocol. The p6 MDA amplified product was purified with the MRC Holland ethanol precipitation method (section 2.3.4.3.1.2.2.) and 3.5 µg of purified product was used in the T7 Endonuclease I digestion reaction. The concentration and purity of the purified T7 Endonuclease I product was measured with the NanoDrop™ 1000 Spectrophotometer and was used to prepare the DNA library for ONT sequencing. The DNA library was purified with AMPure XP beads following ONT's instructions and the concentration of purified library was measured with the Qubit®

2.0 fluorometer. The concentration and purity measurements mentioned in this section are shown in Table 3.9.

Table 3.9: Purification results of the p6 MDA amplified product during ONT's DNA library preparation

	NanoDrop™ 1000 spectrophotometer			**Qubit® 2.0 fluorometer**
Sample name	DNA concentration (ng/ul)	260/280▪	260/230	DNA concentration (ng/ml)
Purified p6 MDA product	621,5	1,90	1,71	
Purified T7 Endonuclease I digested product	19,4	3,23	2,89	
DNA library				<50 Too low to read

▪ Ratio of ~1.8 corresponds to 'pure' DNA and a ratio of ~2.0 corresponds to 'pure' RNA

The recovery of the purified T7 Endonuclease I digested product was lower than the 1 µg input that was required for ONT sequencing when using the Premium whole genome amplification protocol, however, the product was sequenced to determine the efficacy of the protocol and the GridION sequencing device.

3.3.1.4.1. ONT sequencing results for the p6 MDA amplified product

ONT's MinKNOW software showed the real-time sequencing information such as the progression of the sequencing run of the p6 MDA amplified product as mentioned in section 2.3.4.2. Overall, the nanopores were not saturated with a sufficient concentration of the DNA library as noted in Table 3.9. The ONT sequencing run for this sample contained less than 4 000 sequences which is less than what is expected for this sequencing platform. Sequences with a Q score ≥7 were selected and aligned to the HXB2 reference genome. Massive gaps were observed in the alignment and the highest coverage occurred across the 1.2kb region that the p6-PR-RT PCR targeted. Therefore, using a longer PCR extension time did not allow for amplification across the site of integration and HIV integration sites were not recovered with this method.

3.3.2. Amplicons by Ligation

3.3.2.1. Purification of provisionally intact NFL HIV PCR products

Nine provisionally intact NFL HIV-1 patient samples were selected to be sequenced with ONT's GridION. During the time of the project, seven were confirmed to be intact HIV-1 proviral sequences by Illumina® MiSeq™ sequencing and consensus analyses (Katusiime *et al*., 2020). The remaining samples contained deletions and were classified as non-intact for this investigation. The seven intact and two non-intact samples were sequenced by ONT and were included in the downstream consensus construction, sequence analyses and sequencing platform comparisons.

The nine nested NFL amplicons were separated by gel electrophoresis to confirm that the 9kb fragment of interest was present prior to ONT sequencing, these results are depicted in Figure 3.24 and the 9kb fragments of interest are encircled. The first two lanes in the figure contain the non-intact nested NFL products and the remaining lanes contain the seven intact nested NFL products from patient samples. These amplicons were purified with AMPure XP beads (2.3.4.4.1.) and the quantity and quality of the NFL amplicons pre- and post-purification were measured with the NanoDrop™ 1000 Spectrophotometer and are shown in Table 3.10. Each purified 9kb amplicon was processed to prepare a DNA library for ONT sequencing which included end reparation and preparation and the ligation of sequencing adaptors to the ends of the DNA fragment. The DNA libraries were purified with AMPure XP beads (2.3.4.4.3.) and the concentration and purity of each was measured with the NanoDrop™ 1000 Spectrophotometer and these results are included in Table 3.10. This measurement ensured that sufficient DNA library was available for successful ONT sequencing.

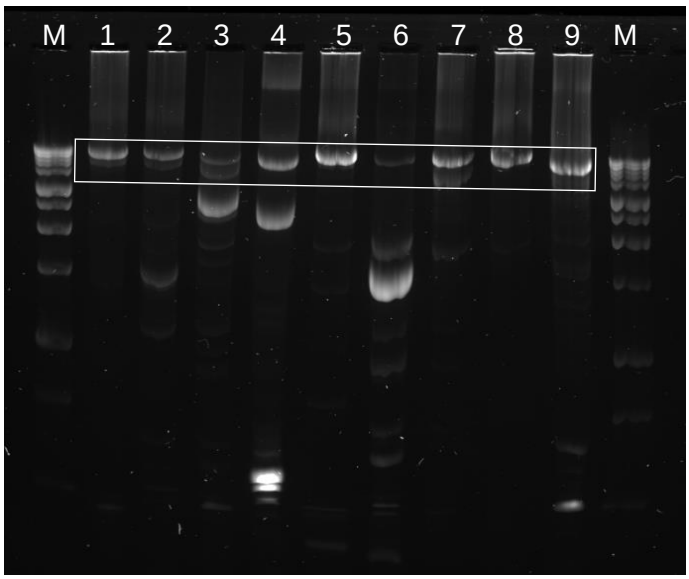

Figure 3.24: The gel electrophoresis separation of nine provisionally intact NFL nested products showing the encircled ~9kb fragments in each lane to be sequenced. *M*, 1kb molecular marker, *1*, 333716 P2D4, *2*, 339606 P3G8, *3*, 339266 P5D4, *4*, 339266 P1C7, *5*, 339266 P1C8, *6*, 339266 P2C7, *7*, 339606 P3D8, *8*, 339606 P3G7, *9*, 340116 P4D1.

Table 3.10: Concentrations of the non-intact and intact NFL amplicons during ONT sequencing preparation

		NFL amplicons			Purified NFL amplicons			DNA library		
Patient ID	**Sample identifier**	DNA concentration (ng/µl)	260/280▪	260/230	DNA concentration (ng/µl)	260/280▪	260/230	DNA concentration (ng/µl)	260/280▪	260/230
333716	P2D4*	476,5	1,74	1,30	131,2	1,90	2,36	-	-	-
339606	P3G8	573,2	1,76	1,34	138,7	1,90	2,36	31,9	1,73	1,42
	P3D8	894,1	1,78	1,25	146,4	1,94	2,35	23,8	1,83	1,36
	P3G7	455,2	1,74	1,24	157,7	1,97	2,37	30,1	1,78	1,35
339266	P5D4	797,9	1,81	1,27	107,7	1,90	2,26	28,3	1,87	1,62
	P1C7	691,2	1,78	1,33	97,5	1,95	2,07	31,6	1,81	1,53
	P1C8	436,1	1,81	1,31	127,9	1,98	2,32	34,3	1,75	1,34
	P2C7	575,2	1,78	1,37	78,2	1,99	2,28	145,9	1,79	1,33
340116	P4D1	571,5	1,78	1,31	125,9	1,94	2,34	49,0	1,53	0,82

* No DNA library sample remaining after purification for measurement

Indicates samples that have been identified as intact HIV-1 sequences

▪ Ratio of ~1.8 corresponds to 'pure' DNA and a ratio of ~2.0 corresponds to 'pure' RNA

3.3.2.2. Analysing ONT sequencing in real-time

The core functions of the MinKNOW software allows for the sequencing run to be monitored in real-time as mentioned in *Step 6* of section 2.3.4.2. and allows for important information about the progression of the sequencing to be immediately relayed. The real-time run information dashboard shows information about the health and status of the nanopores, relays information on the condition of the flow cell (i.e. quality control) and also shows a histogram with the read length distribution. The health and status of the nanopores on the flow cell for a ONT sequencing run is shown in the Channels Panel in Figure 3.25, where each tile represents a single nanopore and bright green tiles correspond to pores that are actively sequencing, dark green shows pores that are available to sequence, dark blue tiles indicate pores that are recovering and light blue tiles are inactive pores. Figure 3.26 shows the total number of fragments of a specific length that have been sequenced and an average fragment length. This fragment length distribution was seen in most of the sequencing experiments with a peak at approximately 9kb. Bimodal peaks were observed in some of the Read Length Histograms, one peak occurred at the expected 9kb fragment length and another occurred between 2kb and 5kb. The sequencing runs for all nine of the NFL samples were stopped once sufficient coverage of the fragment of interest was obtained which equated to 40 - 70 GB of sequencing data for each sample.

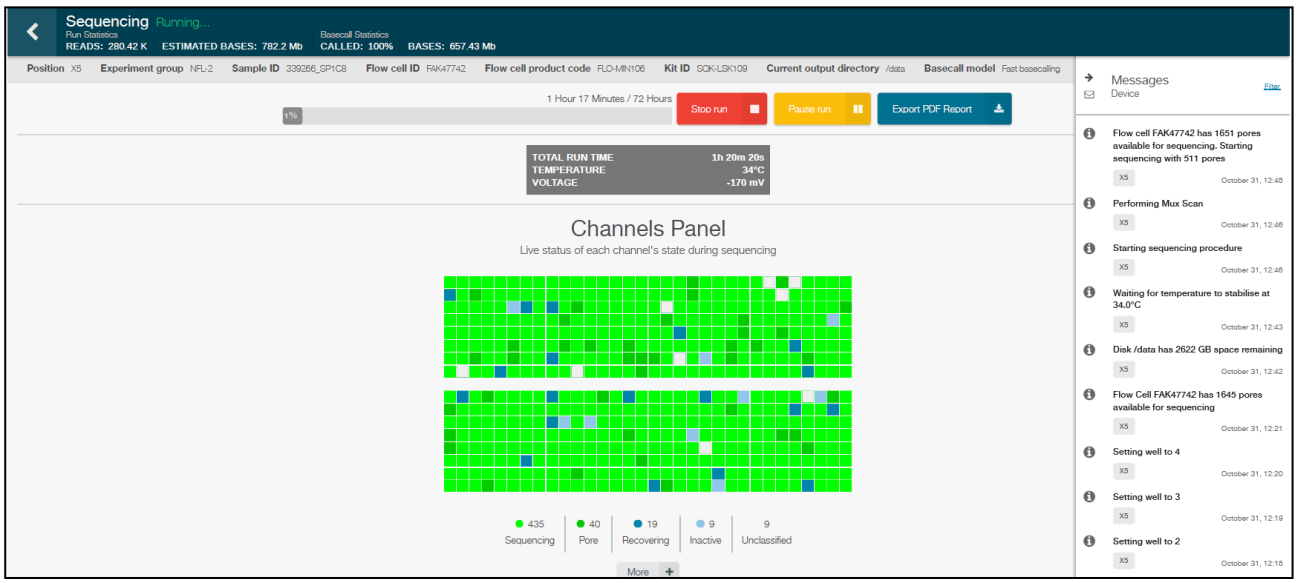

Figure 3.25: ONT real-time dashboard showing the Channels Panel and the status of the nanopores on the flow cell.

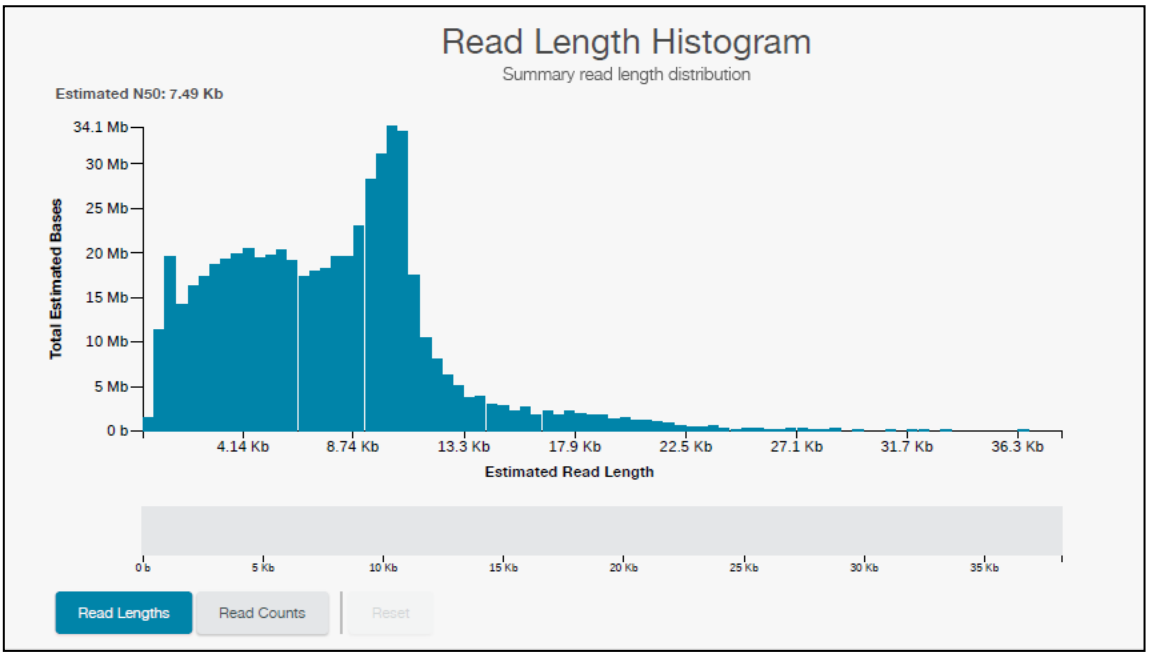

Figure 3.26: ONT real-time dashboard showing a histogram of the read length of the fragments sequenced and the estimated amount of data generated for each read length after 2 hours of sequencing.

3.3.2.3. Phylogenetic analysis and comparison of the Illumina® MiSeq™ and ONT sequencing methods

The developed pipeline was used to construct the consensus sequences from the provisionally intact NFL HIV-1 patient samples that were sequenced by ONT's GridION. The ONT and Illumina® MiSeq™ sequences were aligned using MUSCLE and a tree was generated using maximum parsimony with 5 random rearrangements. The phylogenetic tree is shown in Figure 3.27. The ONT and Illumina® MiSeq™ consensus sequences have been denoted and the sequences belonging to the same patient have been indicated on the righthand side of the figure. The comparison between the two sequencing platforms was further supplemented by Table 3.11 which shows the total aligned similarity percentage between the Illumina® MiSeq™ and ONT consensi of the same patient. Table 3.11 therefore, shows how comparable ONT sequencing is to the golden standard Illumina® MiSeq™. All except one ONT consensus sequence showed a >99% similarity to the Illumina® MiSeq™ sequences. The nucleotide level disagreement between these platforms (<1% difference) was predominantly due to genomic regions with homopolymers (where ONT is known to have a higher read error) and the highly variable HIV-1 *env* gene, where the reads differ considerably from the reference genome.

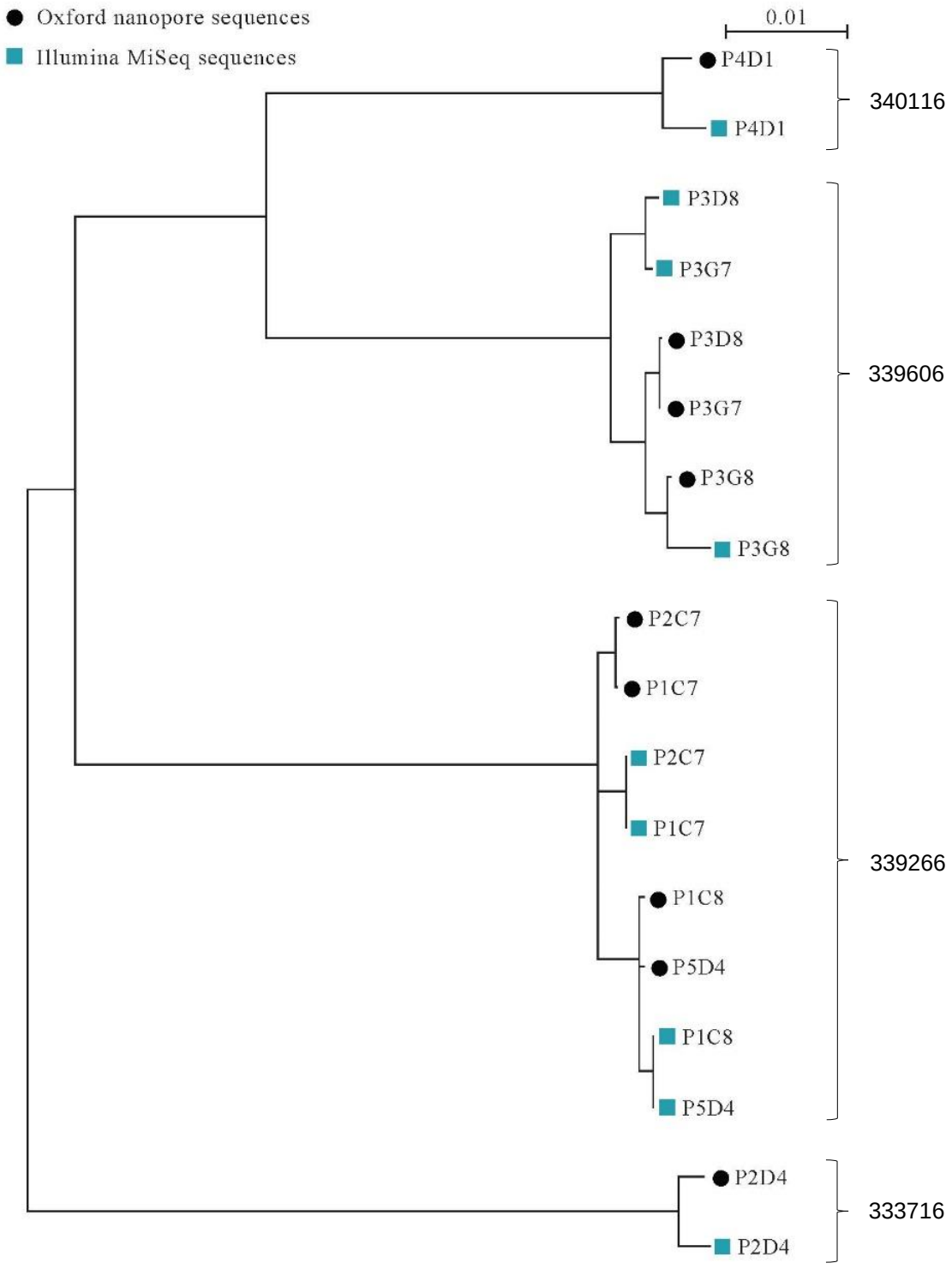

Figure 3.27: Maximum parsimony phylogenetic tree showing the sequencing results of the non-intact and intact NFL HIV-1 patient samples sequenced by ONT and Illumina® MiSeq™.

Table 3.11: Percentage aligned similarity between the ONT consensus and Illumina® MiSeq™ consensus sequences

Patient ID	Sample identifier	Total aligned similarity percentage (%)
333716	P2D4	99.1
339606	P3G8	99.6
339266	P5D4	99.6
	P1C7	99.6
	P1C8	99.7
	P2C7	99.6
339606	P3D8	98.6
	P3G7	99.4
340116	P4D1	99.3

Indicates samples that have been identified as intact HIV-1 sequences

Chapter 4

4 Discussion

4.1. Summary and significance of findings

The purpose of this investigation was to optimise and implement an assay that would detect the integration sites of intact HIV-1 proviral sequences identified in children belonging to the CHER cohort. As intact proviral sequences are rare, first detecting and enriching for them followed by the identification of their respective integration site would be beneficial as a technique like this is not currently available. The adapted ISLA subtype C and ONT's Premium whole genome amplification protocol were utilised to achieve this. Furthermore, with increasing interest in the long-read length and real-time sequencing offered by third-generation sequencing technologies, ONT sequencing was investigated to determine if efficient, real-time and high accuracy sequencing was achievable by utilising ONT's Amplicons by ligation protocol.

4.1.1. ISLA subtype C assay

The main objective of this study was focused on identifying integration sites. The ISLA subtype C assay was utilised as the method for identifying integration sites of intact proviral sequences and yielded unexpected results. The published ISLA method effectively described the identification of integration sites in HIV-1 subtype B patient samples although proviral integrity was not confirmed (Wagner *et al*., 2014) however, in the adapted HIV-1 subtype C version the identification of integration sites was unsuccessful. Various differences can be described between the published ISLA assay and the assay that was optimised in our setting. The most distinct difference is due to the generation of starting material, Wagner *et al* (2014) used two forward primers in *env* and *nef* to generate linearly amplified templates and the adapted ISLA targeted the linearly amplified products in the background of a pre-nested NFL PCR. PCR reactions were optimized for exponential target amplification. However, PCR errors are common and often underappreciated, this includes off-target amplification when primers that are partial complements are extended under low stringency conditions, PCR product recombination, PCR induced mutations and primer-derived products including primer-dimers (Potapov & Ong, 2017, Zanini *et al*., 2017). The 5'-3' exonuclease activity of *Taq* polymerase, could also destroy particular products generated by PCR as it digests products when the enzyme extending a primer encounters another template at a Y-junction which occurs when the product is at least partially

complementary to the template. This could have hampered the recovery of integration sites. Our original expectation was that during PCR cycling, extension of the original template would generate one product per original template per PCR cycle that extends beyond the integration site junction (linear amplification of the original template) shown in Figure 4.1. However, DNA polymerase 5'-3' exonuclease activity could have digested these products. Furthermore, the targeted area of the NFL primer sets undergo exponential amplification and therefore the template containing the integration site is likely present in low concentrations, if it has not been enzymatically digested. Therefore, ISLA assay primers may preferentially bind to the exponentially amplified templates instead as they are present in higher concentrations and consequently integration sites cannot be identified. A further noted difference between the assays is that the original primers were modified to bind to a subtype C HIV-1 genome. The T_m of the primers utilised in the original ISLA and the subtype C ISLA were similar and the published annealing temperatures from Wagner *et al*. (2014) were initially utilised. Due to the small differences in primer T_m, these annealing temperatures may have resulted in mispriming, limiting amplification across the integration site of the original integrated HIV template. For this reason, different annealing temperatures were investigated but integration sites were unrecoverable. In addition to the aforementioned differences, an MDA step was included in the ISLA subtype C assay which was not utilised in the original approach. The MDA step aimed to enrich the linearly amplified intact proviral sequences with the adjacent integration site as they were likely to be present in low concentrations as this would have allowed for a more efficient subtype C ISLA amplification from these templates.

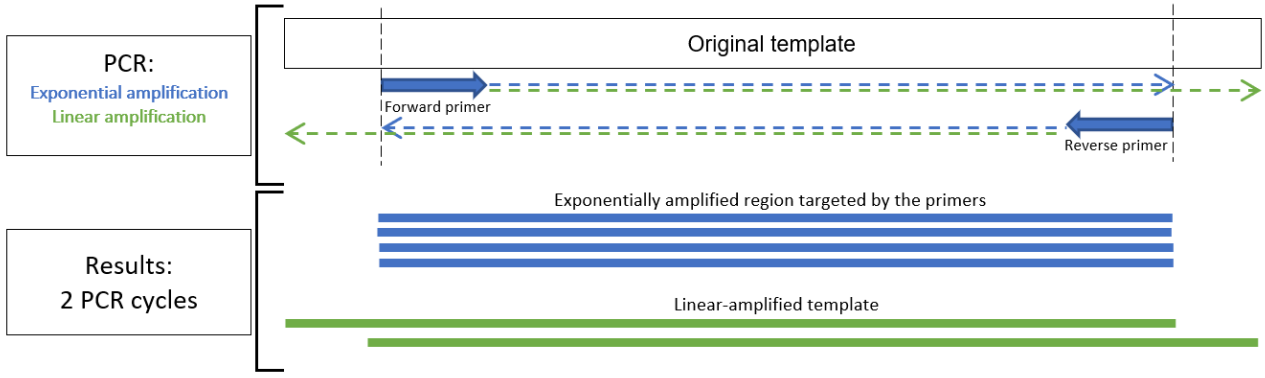

Figure 4.1: Example of the products that are produced during two rounds of PCR. ---> indicates the exponentially amplified region targeted by the primer set. ---> indicates the linearly amplified region where extension from the primer occurs beyond the targeted region.

Although integration sites were not identified through the ISLA subtype C assay, several unique fragments were amplified and sequenced to determine their significance. Twenty-three of the successfully sequenced cloned ISLA samples matched to HIV-1 sequences available on GenBank. Moreover, the majority, equivalent to >39% (9 out of 23), of the sequences matched an intact proviral sequence that was published by Patro and colleagues in 2019. The remaining sequences matched to defective, mutant or RNA HIV-1 sequences of unpublished methods and comparisons between the various methods is therefore not possible. As mentioned previously, the development of assays that allow for the identification of integration sites of intact proviral sequences is currently an important area of research as latently infected cells that harbour intact and potentially replication-competent viruses are important to identify as these cells lead to viral rebound upon cART cessation. Recently, two similar methods employing whole genome amplification of single cells were described by Einkauf *et al*. (2019) and Patro *et al*. (2019) to facilitate these identifications. Therefore, the fact the majority of the sequences in this study returned blastn matches to the clonally expanded intact HIV-1 subtype B proviral sequence from the Patro *et al*. study (2019) is significant as the methods described here and in the published article are somewhat similar. Both studies utilised an NFL PCR on single proviral genomes, although the primer sets were different, to generate a ~9kb HIV-1 template. Secondly, amplification by MDA was used in both studies however, Patro *et al*. (2019) enriched the genomic DNA prior to PCR amplification whereas our study first identified intact HIV-1 proviral sequences and then utilised MDA to enrich for the intact proviral sequences present in the corresponding pre-nested NFL PCR wells. Comparisons between the method described by Einkauf *et al*. (2019), Patro *et al*. (2019) and the subtype C ISLA assay can also be described. Einkauf and colleagues (2019) employed a whole genome amplification approach with MDA on cells with single proviral genomes prior to a NFL amplification step similar to Patro *et al*. (2019) however, the NFL PCR targeted an ≥8kb region of HIV-1. A major similarity between the methods described in this study and those described by Einkauf *et al*. (2019) is that both studies utilised ISLA to identify integration sites. In addition, Einkauf *et al*. (2019) and colleagues confirmed integration sites by utilising two additional integration site assays in conjunction with ISLA, namely, the ligation-mediated PCR (LM-PCR) described by Serrao *et al*. in 2016 and the non-restrictive linear amplification-mediated PCR (nrLAM-PCR) described by Paruzynski *et al*. in 2010. The Patro *et al*. (2019) study investigated five patient samples and identified one expanded clone with an intact proviral sequence which was confirmed to be replication-competent by VOA. Whereas, the Einkauf

et al. (2019) identified eight clonally expanded cells harbouring intact proviral sequences among the three investigated patient samples.

Intact proviral sequences are rare and the laborious steps of recovery have been described in many studies (Bruner *et al*., 2016, Bui *et al*., 2017, Katusiime *et al*., 2020) therefore, identifying the respective integration sites of intact proviruses is more labour intensive, complicated and time-consuming. Few clonally expanded cells harbouring intact proviral sequences have been described to date and although the ISLA subtype C assay approach was unsuccessful, it entailed a novel approach to recover these integration sites in comparison to methods that were recently described. Our approach to identifying integrations sites was unique as MDA is costly to execute on every well present on a 96 well PCR plate, therefore by identifying the wells that contain the intact proviral sequence first and then performing MDA on those wells only, would have significantly reduced costs related to identifying integration sites of intact HIV-1 proviral sequences, if these attempts had been successful.

4.1.2. Identification of HIV-1 integration sites using ONT sequencing

An alternative method for identifying the integration sites of intact HIV-1 proviruses was attempted as ISLA subtype C was unsuccessful, however, integration sites were also unrecoverable using the Premium whole genome amplification method from ONT. The method was attempted as published but significant losses in concentration of both the 8E5 NFL and p6 MDA amplified products were observed and therefore, several other methods of purification were investigated. Although the MRC Holland Ethanol precipitation with the overnight - 20°C incubation method resulted in the best concentration and purity of the T7 Endonuclease I digested product, it did not yield a sufficient concentration to proceed with ONT sequencing as shown in Table 3.7. in the *Results* section. A comparison of the 8E5 NFL MDA amplified product and the patient NFL MDA amplified products revealed that the former product could not be visualised by gel electrophoresis separation or plate viewing whereas the patient NFL MDA amplified products were visualised through both methods. Further comparisons showed that the spectrophotometric measurements of the 8E5 NFL- and patient MDA amplified products measured as DNA. Taken together, the initial thoughts were that the 8E5 NFL MDA amplified product was not double-stranded but rather single-stranded as the EZ Vision™ DNA Dye would have bound to any double-stranded DNA present. Moreover, ONT devices are unable to sequence single-stranded DNA as the

sequencing adaptors, which contain important proteins that facilitate the binding to the sensor chip and sequencing of the fragment via the nanopore, will not ligate to the ends of these fragments. The p6 MDA amplified product which was also not visualised through the plate viewing method but was sequenced through ONT in an attempt to determine whether a low DNA library concentration could be sequenced and potentially resolve some of the questions concerning the structure of the non-visible MDA amplified fragment. The sequencing results of the p6 MDA amplified product revealed that the alignment to the HXB2 consensus sequence contained numerous gaps and deletions which may have occurred as a result of the over frequent cutting of the fragment by the T7 Endonuclease I enzyme thereby resulting in the sequencing of very short fragments which complicate genome assembly. The unexpected cutting carried out by the T7 Endonuclease I enzyme has been hypothesised by many other ONT users although the method established by ONT has been successfully used in two recently published studies in different disciplines (Piñar *et al*., 2020, Harrop *et al*., 2020). Other potential problems that could have impeded successful sequencing by ONT may have occurred as MDA was utilised to enrich the intact proviral sequences and their integration site from a pre-nested NFL PCR product. As these templates are present in low concentrations sufficient enrichment may not have occurred as was described in the ISLA subtype C section above.

4.1.3. Sequencing amplicons with ONT

The major finding in this investigation was that ONT sequencing of products prepared with the Amplicons by Ligation sequencing protocol showed that ONT was an efficient sequencing technology that delivered long read lengths in real-time and generated sequences with acceptable read accuracy which can be used to construct accurate consensus sequences, due to sufficient coverage. ONT reports that a raw read accuracy of 95% can be achieved with the broadly used R9.4.1 nanopores, which were used in this study and that these nanopores have shown the rapid generation of high-quality sequencing data and testing across multiple disciplines. The overall percentage similarity between the consensus sequences generated independently by ONT and Illumina® MiSeq™ was calculated at >99% for 8 of the 9 non-intact/intact HIV-1 proviral sequences. Illumina® MiSeq™ is considered the gold standard sequencing platform and has an average sequencing error rate of between 0.24% and 0.06% per base (Pfeiffer *et al*., 2018). Despite the higher read error of ONT (Weirather *et al*., 2017, Quainoo *et al*., 2017) agreement of ONT and Illumina® MiSeq™ consensus sequences were almost perfect except for HIV-1

env. Earlier studies reported higher error rates for ONT sequencing, a study in 2017 by Weirather *et al*. compared the error rates of PacBio® and ONT and reported error rates of 13.217%, and 12.669%, respectively. By 2018, a study conducted by Jain and colleagues reported that the raw base-called error rate for ONT sequences decreased to <5%. Therefore, ONT's sequencing technology has evolved and the quality of the sequencing data that is generated has significantly improved since its commercial release in 2014. The 1% error of ONT consensus compared to reference Illumina® MiSeq™ consensus reported in this study, may be attributed to the following; 1) regions containing homopolymer stretches were poorly resolved by the R9.4.1 nanopores and basecalling algorithm which often cause deletions within those regions and therefore misalignments to reference sequences, 2) sequences of acceptable quality were mapped to HIV-1 subtype C reference, the resulting coverage and mapping accuracy was lowest for the highly variable *env* region of HIV-1 therefore resolving this region was complicated and 3) the original nested NFL PCR products used to identify intact proviruses by Illumina® MiSeq™ sequencing were depleted, therefore the pre-nested NFL PCR wells corresponding to the intact nested wells were amplified with the nested NFL primer set to generate products for ONT sequencing. Although a high-fidelity enzyme is utilised in the pre- and nested PCR reactions, some errors may have been introduced which could also contribute to the slight differences observed between the Illumina® MiSeq™ and ONT consensi. Many studies that have utilised ONT sequencing have been published, however, at the time of writing only two studies have used ONT sequencing to conduct research on HIV and this may be due to the varying reports on ONT read accuracy which are influenced by poor resolution of homopolymer and highly variable regions. As HIV is a virus that rapidly mutates it's important to utilise a sequencing platform that can accurately detect these mutations. The sequence accuracy from this study could in future be further improved upon by utilising the more accurate ONT R10.3 pores and by improving the HIV *env* alignment algorithms. Nevertheless, this study shows that ONT provides an accurate and fast option for HIV consensus sequence generation, enabled by the long-read lengths, which facilitate accurate alignment and deep sequence coverage which compensates for read error.

4.2. Strengths and limitations of the study

4.2.1. Strengths

Several key strengths were identified in this study. The strengths of the ISLA subtype C assay are limited as many obstacles were encountered throughout the optimisation of the

assay. However, significant advances in technical experience and critical thinking were required to troubleshoot and potentially resolve the problems that were encountered. The majority of the key strengths of this investigation were identified in ONT sequencing. This third-generation sequencing technology allowed for the effortless sequencing of long fragments in real-time and a high coverage across the amplified ~9kb HIV-1 genome was achieved which allowed for accurate consensus sequences to be generated. Despite the higher ONT read error, the approach to obtain consensus sequences provided a very good agreement and comparable accuracy to that obtained with Illumina® MiSeq™. A further strength is that ONT does not require fragmentation and genome assembly is, therefore, less complex and faster. Furthermore, once sufficient coverage is achieved the sequencing run can be terminated, and the remaining resources can be reallocated which reduces the cost of sequencing. ONT is an in-development technology and the technology significantly improves in short periods of time with user input which is provided through the nanopore community platform. ONT often updates and releases new software and chemistry based on the requirements of their users to improve sequencing quality and accuracy. Multiple steps in ONT's protocols can be altered to improve the quality of the sequencing data and includes changes to AMPure XP bead ratio in the purification steps and the ratio of the sequencing adaptors added during the adaptor ligation step. As ONT is a user-orientated sequencing technology any problems that are encountered are quickly resolved by their team of experts.

4.2.2. Limitations

A number of limitations were encountered in this study. Firstly, the success of the ISLA subtype C assay could only be determined after receiving the sequencing data as there are no checkpoint steps between the start of ISLA and the Sanger sequencing. Through discussions with the group that established ISLA (Wagner *et al*., 2014), it became clear that the formation of the loop is generated under very specific conditions that need to be strictly adhered to and any deviation will result in no loop formation and therefore no recovery of integration sites. It is possible that the adjustments made to the original ISLA interfered with the loop formation step. Limitations that were encountered with ONT sequencing included the fact that a high concentration of pure DNA was required for successful high coverage sequencing whereas particular second-generation sequencing platforms such as Illumina® require very low DNA inputs. Low concentration DNA libraries can be sequenced by ONT but the data often contains gaps, gives low coverage and results in the rapid exhaustion of

the nanopores which has a negative impact on the cost of sequencing as flow cells will be discarded after a single use. Efficient bioinformatic pipelines are not as readily available for ONT sequencing as they are for second-generation sequencing technologies, hence the inclusion of a customised pipeline in this study. Particular to ONT is the iterative nature of improvements, which is both a strength and limitation. As the platform constantly improves there is a rapid expansion of research applications. Therefore, custom bioinformatics pipelines for a particular application such as HIV near-full-length sequencing had to be developed as they were not available. Updates to the MinKNOW software are continuously made which improves sequencing accuracy but may lead to delays in sequencing as some updates result in the device not working, this problem was encountered during this study. Library preparation protocols are regularly updated and new protocols are also included however, a protocol that is best suited to your needs may not be available yet and other protocols may require optimisation.

4.3. Investigations and considerations of future studies

The two methods that recently described the simultaneous identification of integration sites and proviral integrity utilise a whole genome amplification approach and use at least 3 different amplification techniques for confirmation, making both approaches expensive, laborious and time-consuming (Einkauf *et al*., 2019, Patro *et al*., 2019). The ISLA subtype C assay and the Premium whole genome amplification approach described in this study was utilised in an attempt to eliminate the increased implications on cost, labour and time. Therefore, future studies should investigate efficient methods that allow for the direct identification of integration sites belonging to intact proviral sequences of HIV-1. Furthermore, third-generation sequencing technologies such as PacBio® and ONT were designed to overcome the limitations present in second-generation sequencing, such as the relative short read-length and long sequencing turnaround time, and facilitate the progression towards faster, cheaper sequencing and potentially efficient identification of integration sites. The initial error rates reported for these technologies were not comparable to those of second-generation sequencing technologies however, recent reports have suggested otherwise and this study further confirms that ONT sequencing, when accompanied by a bioinformatics pipeline that corrects for sequencing error during consensus assembly offers an efficient, rapid and reliable third-generation sequencing technology. Future studies should investigate the error rate of the recently released R10.3 nanopores as they have been specifically designed to overcome the errors encountered

when sequencing homopolymer regions with the R9.4.1 nanopores which were used in this study. The ability to barcode samples with ONT sequencing allows for multiple samples to be sequenced simultaneously, it is also possible to interrupt runs once sufficient coverage is attained with a 'read until' approach, which allows for reuse of flow cells, thereby reducing costs. In future, the effect of barcoding multiple HIV-1 samples and the ability to effectively resolve highly variable regions, when using R10.3 nanopores should be investigated as a lower overall coverage per a sample is achieved when multiple samples are sequenced.

4.4. Conclusion

The integration sites of intact HIV-1 proviral sequences could not be recovered with either of the approaches that were attempted in this study, which included; the adapted subtype C ISLA assay and the altered ONT Premium whole genome amplification protocol. The recovery of integration sites from rare intact HIV-1 proviruses is difficult and cumbersome as has been described here and in other investigations. Assays investigating the clonal expansion of intact proviruses require further development and optimisation to ensure efficient, cost- and time-effective methods are established. As these investigations will allow for a better understanding of how to eliminate the latent reservoir of HIV-1 infected cells harbouring intact provirus, responsible for persistent HIV-1 infection, without disrupting the normal immune functions of CD4+ T cells.

The comparison of ONT and Illumina® MiSeq™ consensi of the same patient samples showed an overall concordance of >99% except for one sample. Furthermore, as described in the *Discussion* section, it was bioinformatically challenging to align reads to the highly variable *env* region of the HIV-1 genome, which effectively reduced coverage and read accuracy for this region. This, together with the poor resolution of homopolymer regions and sequencing of reamplified pre-nested products may account for differences between the ONT and Illumina® MiSeq™ consensi. Overall, ONT's third-generation sequencing technology was able to provide high coverage and accurate consensus sequence generation of the ~9kb intact and non-intact HIV-1 proviral fragments amplified by an NFL PCR. ONT sequencing should be utilised more frequently in research-based settings as it also allows for efficient, low cost and real-time sequencing and data analysis which is extremely useful as many of the currently preferred sequencing technologies have long run times, high reagent costs and often incur longer periods prior to sequence analysis. At the time of writing, only two published studies have utilised ONT for HIV-1 related research and

the findings provided in this study encourage further exploration and use of this sequencing technology especially in research areas that require fast confirmatory results or in studies where high coverage is required to construct accurate consensus sequences.

Reference list

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. & Gouil, Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology,* 21(1):30. doi:10.1186/s13059-020-1935-5.

Arts, E.J. & Hazuda, D.J. 2012. HIV-1 antiretroviral drug therapy. *Cold Spring Harbor Perspectives in Medicine,* 2(4):a007161. doi:10.1101/cshperspect.a007161.

Brown, C. 2019. *Nanopore Community Meeting 2019 technology update. Oxford Nanopore Technologies, Resource Centre.* [Online]. Available: https://nanoporetech.com/resource-centre/nanopore-community-meeting-2019-technology-update [2020, 15 July].

Brown, C. & Buffo, M.J. 2007. Separation and isolation of peripheral blood mononuclear cells in 1ml. *Standard operating procedure. University of Pittsburgh Cancer Institute. Immunologic Monitoring and Cellular products laboratory. SOP-0222*. 2:1-9.

Bruner, K.M., Murray, A.J., Pollack, R.A., Soliman, M.G., Laskey, S.B., Capoferri, A.A., *et al*. 2016. Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nature Medicine,* 22:1043–1049. doi:10.1038/nm.4156.

Buermans, H.P. & den Dunnen, J.T. 2014. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta,* 1842(10):1932- 1941. doi:10 .1016/j.bbadis.2014.06.015.

Bui, J.K., Sobolewski, M.D., Keele, B.F., Spindler, J., Musick, A., Wiegand, A., *et al*. 2017. Proviruses with identical sequences comprise a large fraction of the replication-competent HIV reservoir. *PLoS Pathogens,* 13:e1006283. doi:10.1371/journal.ppat.100 6283.

Bunders, M.J., van der Loos, C.M., Klarenbeek, P.L., van Hamme, J.L., Boer, K., Wilde, J.C.H., *et al*. 2012. Memory CD4(+)CCR5(+) T cells are abundantly present in the gut of newborn infants to facilitate mother-to-child transmission of HIV-1. *Blood,* 120(22):4383-4390. doi:10.1182/blood-2012-06-437566.

Buzon, M.J., Martin-Gayo, E., Pereyra, F., Ouyang, Z., Sun, H., Li, J.Z., *et al*. 2014. Long-term antiretroviral treatment initiated at primary HIV-1 infection affects the size,

composition, and decay kinetics of the reservoir of HIV-1-infected CD4 T cells. *Journal of Virology,* 88(17):10056-10065. doi:10.1128/JVI.01046-14.

Coffin, J. & Swanstrom, R. 2013. HIV pathogenesis: dynamics and genetics of viral populations and infected cells. *Cold Spring Harbor Perspectives in Medicine*, 3(1):a012526. doi:10.1101/cshperspect.a012526.

Cohn, L.B., Silva, I.T., Oliveira, T.Y., Rosales, R.A., Parrish, E.H., Learn, G.H., *et al*. 2015. HIV-1 integration landscape during latent and active infection. *Cell,* 160(3):420-32. doi:10.1016/j.cell.2015.01.020.

Cotton, M.F., Violari, A., Otwombe, K., Panchia, R., Dobbels, E., Rabie, H., *et al*. 2013. Early time-limited antiretroviral therapy versus deferred therapy in South African infants infected with HIV: results from the children with HIV early antiretroviral (CHER) randomised trial. *Lancet,* 382(9904):1555-1563. doi:10.1016/S0140-6736(13)61409-9.

Das, A.T., Pasternak, A.O. & Berkhout, B. 2019. On the generation of the MSD-Ψ class of defective HIV proviruses. *Retrovirology,* 16(1):19. doi:10.1186/s12977-019-0481-2.

De Maio, N., Shaw, L.P., Hubbard, A., George, S., Sanderson, N.D., Swann, J., *et al*. 2019. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics,* 5(9):e000294. doi:10.1099/mgen.0.000294.

Einkauf, K.B., Lee, G.Q., Gao, C., Sharaf, R., Sun, X., Hua, S., *et al*. 2019. Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *The Journal of Clinical Investigation,* 129(3):988-998. doi:10.1172/JCI124291.

Engelman, A. & Cherepanov, P. 2013. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nature Reviews. Microbiology,* 10(4):279– 290. doi:10.1038/nrmicro2747.

Folks, T.M., Powell, D., Lightfoote, M., Koenig, S., Fauci, A.S., Benn, S., *et al*. 1986. Biological and biochemical characterization of a cloned Leu-3- cell surviving infection with the acquired immune deficiency syndrome retrovirus. *The Journal of Experimental Medicine,* 164 (1): 280–290. doi:10.1084/jem.164.1.280.

Gartner, M.J., Roche, M., Churchill, M.J., Gorry, P.R. & Flynn, J.K. 2020. Understanding the mechanisms driving the spread of subtype C HIV-1. *EBioMedicine*, 53:102682. doi:10.1016/j.ebiom.2020.102682.

German Advisory Committee Blood (Arbeitskreis Blut), Subgroup 'Assessment of Pathogens Transmissible by Blood'. 2016. Human Immunodeficiency Virus (HIV). *Transfusion medicine and hemotherapy*, *43*(3):203–222. doi:10.1159/000445852.

Göpfrich, K. & Judge, K. 2018. *Science in School. Decoding DNA with a pocket-sized sequencer*. [Online]. Available: https://www.scienceinschool.org/content/decoding-dna-pocket-sized-sequencer [2020, 20 May].

Günthard, H.F. & Scherrer, A.U. 2016. HIV-1 Subtype C, Tenofovir, and the Relationship With Treatment Failure and Drug Resistance. *The Journal of Infectious Diseases,* 214(9):1289-1291. doi:10.1093/infdis/jiw214.

Han, Y., Lassen, K., Monie, D., Sedaghat, A.R., Shimoji, S., Liu, X., *et al*. 2004. Resting CD4+ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. *Journal of virology*, *78*(12), 6122–6133. doi:10.1128/JVI.78.12.6122-6133.2004.

Harrop, T.W.R., Le Lec, M.F., Jauregui, R., Taylor, S.E., Inwood, S.N., van Stijn, T., *et al*. 2020. Genetic Diversity in Invasive Populations of Argentine Stem Weevil Associated with Adaptation to Biocontrol. *Insects,* 11:441. doi:10.3390/insects11070441.

Health Systems Trust. 2016. The 90-90-90 Compendium: An Introduction to 90-90-90 in South Africa. *Health Systems Trust,* 1:1-35.

Ho, Y.C., Shan, L., Hosmane, N.N., Wang, J., Laskey, S.B., Rosenbloom, D.I.S., *et al*. 2013. Replication-Competent Noninduced Proviruses in the Latent Reservoir Increase Barrier to HIV-1 Cure. *Cell,* 155(3):540–551. doi:10.1016/j.cell.2013.09.020.

Hong, F., Aga, E., Cillo, A.R., Yates, A.L., Besson, G., Fyne, E., *et al*. 2016. Novel Assays for Measurement of Total Cell-Associated HIV-1 DNA and RNA. *Journal of Clinical Microbiology,* 54:902–911. doi:10.1128/JCM.02904-15.

Hosmane, N.N., Kwon, K.J., Bruner, K.M., Capoferri, A.A., Beg, S., Rosenbloom, D.I.S., *et al*. 2017. Proliferation of latently infected CD4+ T cells carrying replication-competent HIV-1: potential role in latent reservoir dynamics. *The Journal of Experimental Medicine,* 214:959–972. doi:10.1084/jem.20170193.

Illumina®. 2017. *An introduction to Next-Generation Sequencing Technology*. [Online]. Available: https://emea.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf [2020, 7 June].

Ion Proton vs Illumina HiSeq2500 vs SOLiD 5500. 2013. [Online]. Available: http://www.innofoodsee.eu/downloads/compare.pdf [2020, 10 July].

Katusiime, M.G., Halvas, E.K., Wright, I., Joseph, K., Bale, M.J., Kirby-McCullough, B., *et al*. 2020. Intact HIV Proviruses Persist in Children Seven to Nine Years after Initiation of Antiretroviral Therapy in the First Year of Life. *Journal of Virology,* 94(4)e01519-19.

Kogan, M. & Rappaport, J. 2011. HIV-1 accessory protein Vpr: relevance in the pathogenesis of HIV and potential for therapeutic intervention. *Retrovirology,* 8:25. doi:10.1186/1742-4690-8-25.

Krishnakumar, R., Sinha, A., Bird, S.W., Jayamohan, H., Edwards, H.S., Schoeniger, J.S., *et al*. 2018. Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. *Scientific Reports,* 8(1):3159. doi:10.1038/s41598-018-21484-w.

Kuhn, L., Paximadis, M., Da Costa Dias, B., Loubser, S., Strehlau, R., Patel, F., *et al*. 2018. Age at antiretroviral therapy initiation and cell-associated HIV-1 DNA levels in HIV-1-infected children. *PLoS One*, 13(4):e0195514. doi:10.1371/journal.pone.0195514.

Laskey, S.B., Pohlmeyer, C.W., Bruner, K.M. & Siliciano, R.F. 2016. Evaluating Clonal Expansion of HIV-Infected Cells: Optimization of PCR Strategies to Predict Clonality. *PLoS Pathogens,* 12(8):e1005689.

Le, T., Wright, E.J., Smith, D.M., He, W., Catano, G., Okulicz, J.F., *et al*. 2013. Enhanced CD4+ T-cell recovery with earlier HIV-1 antiretroviral therapy. *The New England Journal of Medicine*, 368(3):218-230. doi:10.1056/NEJMoa1110187.

Leitner, T., Korber, B., Daniels, M., Calef, C. & Foley, B. 2005. HIV-1 subtype and circulating recombinant form (CRF) reference sequences, 2005. *HIV sequence compendium*, *2005*, pp.41-48.

Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics,* 34(18):3094–3100. doi:10.1093/bioinformatics/bty191.

Li, G. & De Clercq, E. 2016. HIV Genome-Wide Protein Associations: a Review of 30 Years of Research. *Microbiology and Molecular Biology Reviews,* 80(3):679- 731. doi:10.1128/MMBR.00065-15.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al*. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25(16):2078- 2079. doi:10.1093/bioinformatics/btp352.

Li, L., Liang, S., Chen, L., Liu, W., Li, H., Liu, Y., *et al*. 2010. Genetic characterization of 13 subtype CRF01_AE near full-length genomes in Guangxi, China. *AIDS Research and Human Retroviruses,* 26(6):699-704. doi:10.1089/aid.2010.0026.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., *et al*. 2012. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine & Biotechnology,* 2012:251364. doi:10.1155/2012/251364.

Liu, R., Simonetti, F.R. & Ho, Y. 2020. The forces driving clonal expansion of the HIV-1 latent reservoir. *Virology Journal*, 17(1):4. doi:10.1186/s12985-019-1276-8.

Loman, N.J., Quick, J. & Simpson, J.T. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12:733–735. doi:10.1038/nmeth.3444.

Lorenzi, J.C., Cohen, Y.Z., Cohn, L.B., Kreider, E.F., Barton, J.P., Learn, G.H., *et al*. 2016. Paired quantitative and qualitative assessment of the replication-competent HIV-1 reservoir and comparison with integrated proviral DNA. *PNAS,* 113(49):E7908– E7916. doi:10.1073/pnas.1617789113.

Lorenzo-Redondo, R., Fryer, H.R., Bedford, T., Kim, E., Archer, J., Pond, S.L.K., *et al*. 2016. Persistent HIV-1 replication maintains the tissue reservoir during therapy. *Nature,* 530(7588):51-56. doi:10.1038/nature16933.

Lu, H., Giordano, F. & Ning, Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics & Bioinformatics,* 14:265-279. doi:10.1016/j.gpb.2016.05.004.

Maldarelli, F., Wu, X., Su, L., Simonetti, F.R., Shao, W., Hill, S., *et al*. 2014. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science,* 345(6193):179-83. doi:10.1126/science.1254194.

Martinez, D.R., Permar, S.R. & Fouda, G.G. 2015. Contrasting adult and infant immune responses to HIV infection and vaccination. *Clinical and Vaccine Immunology,* 23(2):84– 94. doi:10.1128/CVI.00565-15.

Murray, A.J., Kwon, K.J., Farber, D.L. & Siliciano, R.F. 2016. The Latent Reservoir for HIV-1: How Immunologic Memory and Clonal Expansion Contribute to HIV-1 Persistence. *Journal of Immunology,* 197(2):407-417. doi:10.4049/jimmunol.1600343.

NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. 2016. *Nucleic Acids Research,* 44(D1):D7-D19. doi:10.1093/nar/gkv1290.

Nkeze, J., Li, L., Benko, Z., Li, G. & Zhao, R.Y. 2015. Molecular characterization of HIV-1 genome in fission yeast *Schizosaccharomyces pombe*. *Cell & Bioscience,* 5:47. doi:10.1186/s13578-015-0037-7.Okoye, A.A. & Picker, L.J. 2013. CD4+ T cell depletion in HIV infection: mechanisms of immunological failure. *Immunological Reviews,* 254(1):54–64. doi:10.1111/imr.12066.

Olabode, A.S., Avino, M., Ng, G.T., Abu-Sardanah, F., Dick, D.W. & Poon, A.F.Y. 2019. Evidence for a recombinant origin of HIV-1 Group M from genomic variation. *Virus Evolution,* 5(1):vey039. doi:10.1093/ve/vey039.

Oxford Nanopore Technologies. 2018. *Final Product Brochure*. [Online]. Available: https://nanoporetech.com/sites/default/files/s3/Product_brochure_Final_July_2018.pdf [2020, 15 July].

Oxford Nanopore Technologies. 2020. *Nanoporetech. Analysis solutions for nanopore sequencing data*. [Online]. Available: https://nanoporetech.com/nanopore-sequencing-data-analysis [2020, 21 May].

Oxford Nanopore Technologies. 2020. *Nanoporetech. Configure your device package*. [Online]. Available: https://store.nanoporetech.com/configure-minion-basic [2020, 5 September].

Oxford Nanopore Technologies. 2020. *Nanoporetech. How does nanopore DNA/RNA sequencing work?*. [Online]. Available: https://nanoporetech.com/how-it-works [2020, 2 August].

Pacific Biosciences. 2019. *SMRT Science. HiFi Reads for Highly Accurate Long-Read Sequencing*. [Online]. Available: https://www.pacb.com/smrt-science/attachment/how-to-get-hifi-reads_v2/ [2020, 12 July].

Pacific Biosciences. 2020. *SMRT Science. Delivering Highly Accurate Long Reads to Drive Discovery in Life Science*. [Online]. Available: https://www.pacb.com/smrt-science/smrt-sequencing/ [2020, 2 August].

Parikh, U.V., McCormick, K., van Zyl, G.U. & Mellors, J.W. 2017. Future technologies for monitoring HIV drug resistance and cure. *Current Opinion in HIV and AIDS,* 12(2):182–189. doi:10.1097/COH.0000000000000344.

Paruzynski, A., Arens, A., Gabriel, R., Bartholomae, C.C., Scholz, S., Wang, W., *et al*. 2010. Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing. *Nature Protocols,* 5(8):1379-1395. doi:10.1038/nprot.2010.87.

Patro, S.C., Brandt, L.D., Bale, M.J., Halvas, E.K., Joseph, K.W., Shao, W., *et al*. 2019. Combined HIV-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors. *PNAS,* 116(51):25891- 25899. doi:10.1073/pnas.1910334116.

Payne, A., Holmes, N., Rakyan, V. & Loose, M. 2019. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*, 35(13):2193- 2198. doi:10.1093/bioinformatics/bty841.

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J.L., *et al*. 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports,* 8:10950. doi:10.1038/s41598-018-29325-6.

Piñar, G., Poyntner, C., Lopandic, K., Tafer, H. & Sterflinger, K. 2020. Rapid diagnosis of biological colonization in cultural artefacts using the MinION nanopore sequencing technology. *International Biodeterioration & Biodegradation,* 148:104908. doi:10.1016/j.ibiod.2020.104908.

Pinzone, M.R. & O'Doherty, U. 2018. Measuring integrated HIV DNA ex vivo and in vitro provides insights about how reservoirs are formed and maintained. *Retrovirology,* 15(1):22. doi:10.1186/s12977-018-0396-3.

Potapov, V. & Ong, J.L. 2017. Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLoS One,* 12(1):e0169774. doi:10.1371/journal.pone.0169774.

Quainoo, S., Coolen, J.P.M., van Hijum, S.A.F.T., Huynen, M.A., Melchers, W.J.G., van Schaik, W., *et al*. 2017. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clinical Microbiology Reviews,* 30:1015-1063. doi:10.1128/CMR.00016-17.

Quillent, C., Dumey, N., Dauguet, C. & Clavel, F. 1993. Reversion of a polymerase-defective integrated HIV-1 genome. *AIDS Research and Human Retroviruses,* 9(10):1031-1037. doi:10.1089/aid.1993.9.1031.

Reeves, D.B., Duke, E.R., Wagner, T.A., Palmer, S.E., Spivak, A.M. & Schiffer, J.T. 2018. A majority of HIV persistence during antiretroviral therapy is due to infected cell proliferation. *Nature Communications,* 9:4811. doi:10.1038/s41467-018-06843-5.

Rhoads, A. & Au, K.F. 2015. PacBio Sequencing and Its Applications. *Genomics Proteomics & Bioinformatics,* 13:278-289. doi:10.1016/j.gpb.2015.08.002.

Sáez-Cirión, A., Bacchus, C., Hocqueloux, L., Avettand-Fenoel, V., Girault, I., Lecuroux, C., *et al*. 2013. Post-treatment HIV-1 controllers with a long-term virological remission after the interruption of early initiated antiretroviral therapy ANRS VISCONTI Study. *PLoS Pathogens,* 9(3):e1003211. doi:10.1371/journal.ppat.1003211.

Sanger, F., Nicklen, S. & Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *PNAS,* 74(12):5463-5467. doi:10.1073/pnas.74.12.5463.

Schröder, A.R., Shinn, P., Chen, H., Berry, C., Ecker, J.R. & Bushman, F. 2002. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell,* 110(4):521-529. doi:10.1016/s0092-8674(02)00864-4.

Serrao, E., Cherepanov, P. & Engelman, A.N. 2016. Amplification, Next-generation Sequencing, and Genomic DNA Mapping of Retroviral Integration Sites. *Journal of Visualized Experiments,* (109):53840. doi:10.3791/53840.

Shalekoff, S., Gray, G.E. & Tiemessen, C.T. 2004. Age-Related Changes in Expression of CXCR4 and CCR5 on Peripheral Blood Leukocytes from Uninfected Infants Born to Human Immunodeficiency Virus Type 1-Infected Mothers. *Clinical and Diagnostic Laboratory Immunology,* 11(1):229-234. doi:10.1128/CDLI.11.1.229–234.2004.

Siliciano, J.D., Kajdas, J., Finzi, D., Quinn, T.C., Chadwick, K., Margolick, J.B., *et al*. 2003. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting $CD4^{+}$T cells. *Nature Medicine,* 9:727–728. doi:10.1038/nm880.

Simonetti, F.R., Sobolewski, M.D., Fyne, E., Shao, W., Spindler, J., Hattori, J., *et al*. 2016. Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. *PNAS,* 113(7):1883-1888. doi:10.1073/pnas.1522675113.

SPARTAC Trial Investigators., Fidler, S., Porter, K., Ewings, F., Frater, J., Ramjee, G., *et al*. 2013. Short-course antiretroviral therapy in primary HIV infection. *The New England Journal of Medicine*, 368(3):207-217. doi:10.1056/NEJMoa1110039.

Stöhr, W., Fidler, S., McClure, M., Weber, J., Cooper, D., Ramjee, G., *et al*. 2013. Duration of HIV-1 viral suppression on cessation of antiretroviral therapy in primary infection correlates with time on therapy. *PLoS One*, 8(10):e78287. doi:10.1371/journal.pone.0078287.

Taylor, B.S., Sobieszczyk, M.E., McCutchan, F.E. & Hammer, S.M. 2008. The Challenge of HIV-1 Subtype Diversity. The *New England Journal of Medicine*, 358(15):1590–1602. doi:10.1056/NEJMra0706737.

Tobin, N.H. & Aldrovandi, G.M. 2013. Immunology of Pediatric HIV Infection. *Immunological Reviews,* 254(1): 143–169. doi:10.1111/imr.12074.

UNAIDS. 2020. *Global HIV & AIDS statistics. 2020 fact sheet*. [Online]. Available: https://www.unaids.org/en/resources/fact-sheet [2020, July 13].

Van Zyl, G.U., Katusiime, M.G., Wiegand, A., McManus, W.R., Bale, M.J., Halvas, E.K., *et al*. 2017. No evidence of HIV replication in children on antiretroviral therapy. *Journal of Clinical Investigation,* 127(10):3827-3834. doi:10.1172/JCI94582.

Vandergeeten, C., Fromentin, R., Merlini, E., Lawani, M.B., DaFonseca, S., Bakeman, W., *et al*. 2014. Cross-clade ultrasensitive PCR-based assays to measure HIV persistence in large-cohort studies. *Journal of Virology,* 88(21):12385-12396. doi:10.1128/JVI.00609-14.

Vanhamel, J., Bruggemans, A. & Debyser, Z. 2019. Establishment of latent HIV-1 reservoirs: what do we really know?. *Journal of Virus Eradication,* 5(1):3-9.

Violari, A., Cotton, M.F., Gibb, D.M., Babiker, A.G., Steyn, J., Madhi, S.A., Jean-Philippe, P. & McIntyre, J.A. 2008. Early Antiretroviral Therapy and Mortality among HIV-Infected Infants.New England Journal of Medicine. 359(21):2233–2244.

Wagner, T.A., McLaughlin, S., Garg, K., Cheung, C.Y.K., Larsen, B.B., Styrchak, S., *et al*. 2014. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science,* 345(6196):570-3. doi:10.1126/science.1256304.

Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X., *et al*. 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis [version 2; referees:2 approved]. *F1000Research,* 6:100. doi:10.12688/f1000research.10571.2.

Wells, D.W., Guo, S., Shao, W., Bale, M.J., Coffin, J.M., Hughes, S.H., *et al*. 2020. An analytical pipeline for identifying and mapping the integration sites of HIV and other retroviruses. *BMC Genomics,* 21:216. doi:10.1186/s12864-020-6647-4.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P., Hall, R.J., Concepcion, G.T., *et al*. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155-1162. doi:10.1038/s41587-019-0217-9.

Zanini, F., Brodin, J., Albert, J. & Neher, R.A. 2017. Error rates, PCR recombination, and sampling depth in HIV-1 whole genome deep sequencing. *Virus Research,* 239:106-114. doi: 10.1016/j.virusres.2016.12.009.

Zheng, N.N., Kiviat, N.B., Sow, P.S., Hawes, S.E., Wilson, A., Diallo-Agne, H., *et al*. 2004. Comparison of human immunodeficiency virus (HIV)-specific T-cell responses in HIV-1- and HIV-2-infected individuals in Senegal. *Journal of Virology*, 78(24):13934-13942. doi:10.1128/JVI.78.24.13934-13942.2004.

Addendum A

SARS-CoV-2 in Cape Town, South Africa investigation report

Molecular epidemiology and genetic diversity of SARS-CoV-2 in Cape Town, South Africa

Susan Engelbrecht, Kayla Delaney, Bronwyn Kleinhans, Eduan Wilkinson, Houriiyah Tegally, Tania Stander, Gert van Zyl, Wolfgang Preiser, Tulio de Oliveira

Introduction:

SARS-CoV-2 is responsible for the COVID-19 disease and was originally reported during a local pneumonia outbreak of unknown origin in Wuhan City, China in December 2019. COVID-19 rapidly became an emerging pandemic and data reported by WHO on the 6th of September 2020 reported that a total of 26 763 217 cases have been confirmed and has resulted in 876 616 deaths globally [https://covid19.who.int/]. Furthermore, the data provided by WHO shows South Africa is ranked as the country with the 6th highest number of cases globally. As of yet, no vaccines or antiviral agents are available to combat the virus although many are currently in development or testing in clinical trials.

In South Africa, early reports of the epidemic identified Cape Town in the Western Cape province as the epicentre. However, prior to this study the number of SARS-CoV-2 introductions and the source of these introductions into Cape Town was unknown which lead to the following questions: 1) was the SARS-CoV-2 introduction into Cape Town a result of a single or multiple introductory event(s)? and 2) what was the source and timing of these events?

This study aimed to track the start of the epidemic in Cape Town by sequencing forty-seven complete viral genomes from early SARS-CoV-2 samples by utilising third-generation sequencing offered by ONT.

Methods:

SARS-CoV-2 viral RNA was detected in patients by diagnostic real-time PCR, reverse transcription and PCR of the complete viral genome was completed with the PCR tiling method available at https://artic.network/ncov-2019. The amplified products were sequenced using ONT's GridION device. The raw reads were assembled and the consensus sequences were obtained by utilising the bioinformatics pipeline available at

https://www.krisp.org.za/ngs-sa/ngs-sa_network_for_genomic_surveillance_south_Africa. The clades were assigned and quality assurance was performed in Nextclade [https://clades.nextstrain.org/]. Finally, the lineages were assigned using Pangolin (Rambaut *et al.*, 2020)[https://cov-lineages.org/] and all the obtained sequences were deposited in GISAID [https://www.gisaid.org/].

Results:

The demographic information and Ct values of the patients were recorded and overall, lower Ct scores were recorded in younger patients, indicative of higher viral loads, and these samples resulted in higher genome coverage. Three clades, namely, 19A, 20A and 20B were identified by Nextstrain phylogenetic analysis and Nextclade. Each clade is defined by specific designated mutations whereas lineages were assigned by the Pangolin software. A lineage is defined as a cluster of sequences that have been observed in geographically distinct regions in conjunction with evidence of ongoing transmission within that region. Two unique variant mutations were identified in Cape Town. A5209G represented a new lineage with a specific amino acid mutation and A24862G was present at a much higher frequency in the Cape Town sequences in comparison to their global distribution. Overall, the clade and lineage data show that multiple introductory events of SARS-CoV-2 were responsible for the introduction into Cape Town. The data further indicates that all the introductions into Cape Town occurred between the last week of February and within the first two weeks of March 2020 prior to the enforcement of strict lockdown regulations in South Africa.

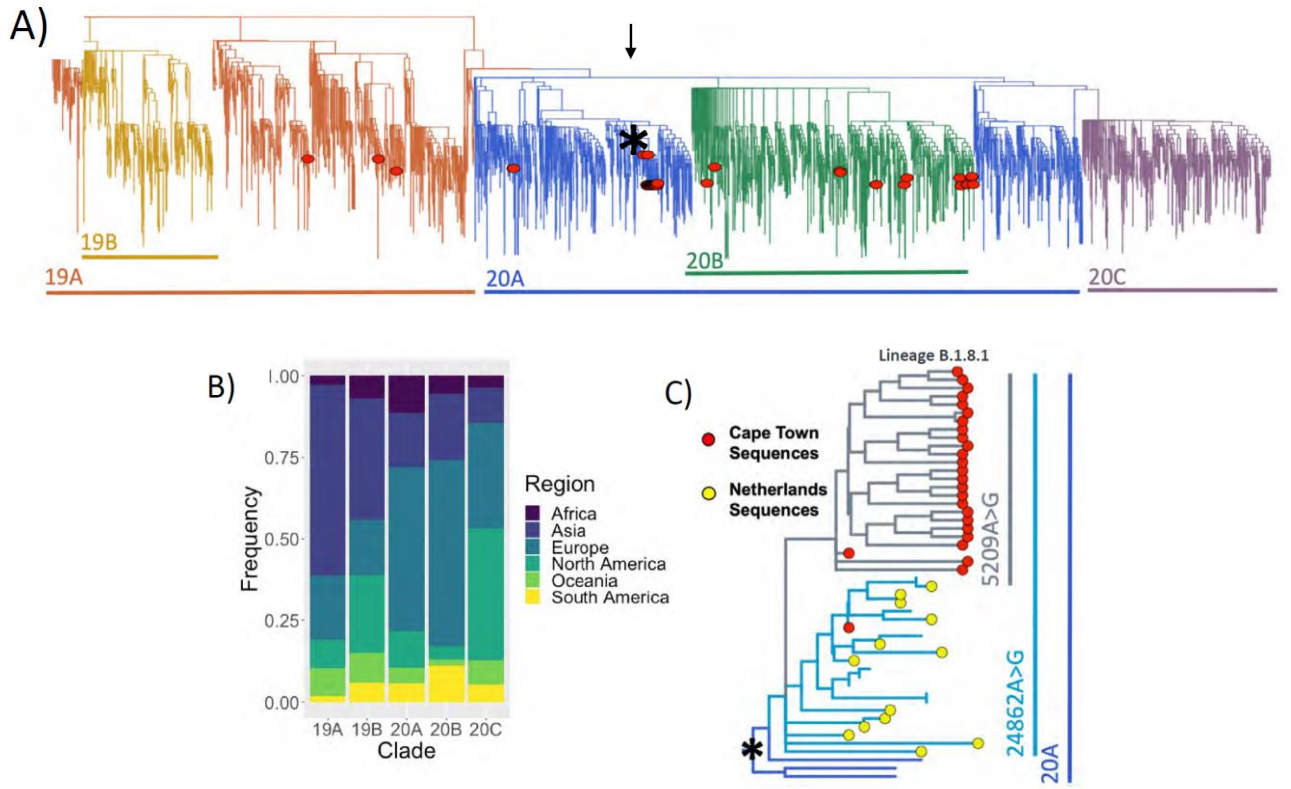

Figure A1: Phylogenetic and clade analysis of the Cape Town Western Cape, South Africa SARS-CoV-2 sequences. (A) A time-scaled Maximum likelihood tree showing a total of 3620 sequences together with the 47 genotypes obtained from Cape Town with the major clades of SARS-CoV-2 indicated with by ●. (B) Stacked bar plot showing the clade frequencies and distributions by region, clade 20A (where most of the Cape Town sequences lie) is over-represented and shows origins from Europe. (C) Monophyletic cluster indicated by * of the Cape Town sequences. This cluster shows the closest divergence to isolates that originated in the Netherlands and the emergence and presence of the unique Cape Town lineage and mutation (5209A>G) is shown within the cluster.

Conclusion:

The presence of clades 19A, 20A and 20B were observed through the genomic and epidemiological data which provided unique insights into the spread and transmission of the early SARS-CoV-2 epidemic in Cape Town, South Africa. Molecular clock analyses confirmed that all the introductions into Cape Town occurred within 3 weeks starting in the last week of February 2020, prior to lockdown in March 2020. One of the introductions, first identified in this study, diversified by acquiring the A5209G mutation prior to causing a large cluster outbreak. Observing and following the acquisition of mutations will allow for a better understanding of how the virus spreads between different locations within a city and throughout South Africa.

Reference list:

Artic Network. 2020. *SARS-CoV-2.* [Online]. Available: https://artic.network/ncov-2019 [2020, 6 September].

Network for Genomic Surveillance in South Africa. 2020. *NGS-SA: Network for genomic surveillance in South Africa.* [Online]. Available: https://www.krisp.org.za/ngs-sa/ngs-sa_network_for_genomic_surveillance_south_Africa [2020, 6 September].

Nextclade. 2020. *Nextclade. Clade assignment, mutation calling, and sequence quality checks.* [Online]. Available: https://clades.nextstrain.org/ [2020, 6 September].

Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., *et al*. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology.* doi:10.1038/s41564-020-0770-5.

World Health Organisation. 2020. *WHO Coronavirus Disease (COVID-19) Dashboard.* [Online]. Available: https://covid19.who.int/ [2020, 6 September].

Addendum B

Average Read Length Histograms for the seven intact HIV-1 proviral sequences using ONT's Amplicons by Ligation protocol for sequencing. The two non-intact samples are excluded as they were sequenced simultaneously (on separate flow cells) and individual Read Length Histograms were not available when using these sequencing settings.

All histograms are based on the average read lengths that were sequenced within the first hour of the ONT sequencing run except for sample 339266 P5D4 which was taken 5 minutes into the sequencing run.

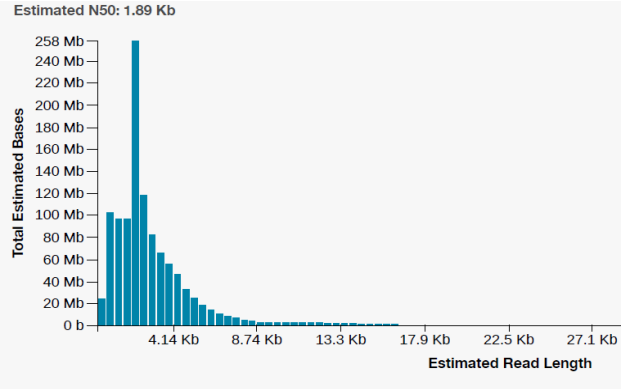

Figure B1: Read length histogram for sample 339266 P2C7

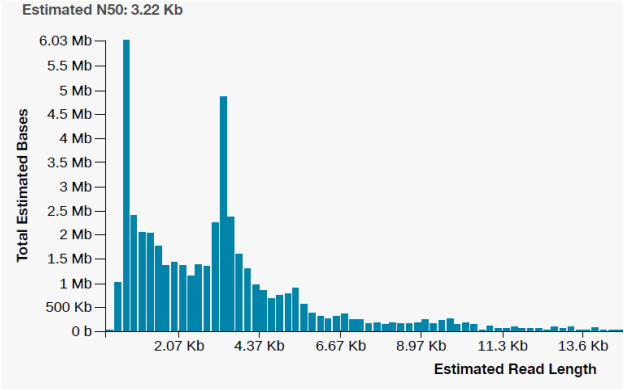

Figure B2: Read length histogram for sample 339266 P5D4

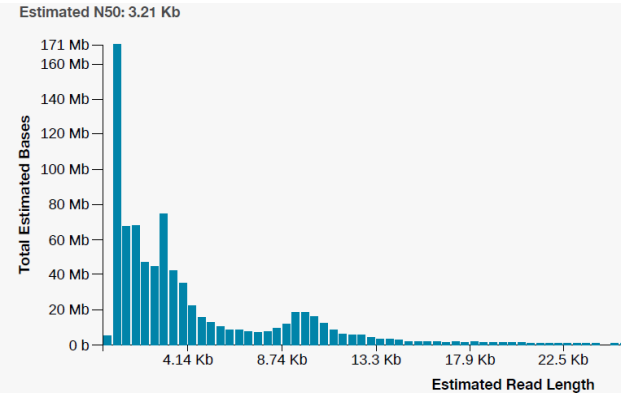

Figure B3: Read length histogram for sample 339266 P1C7

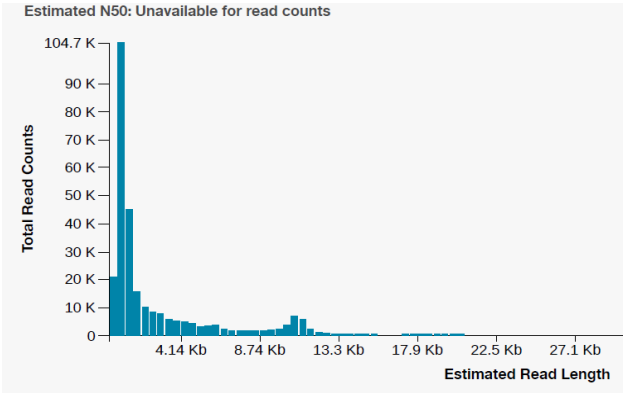

Figure B4: Read length histogram for sample 339266 P1C8

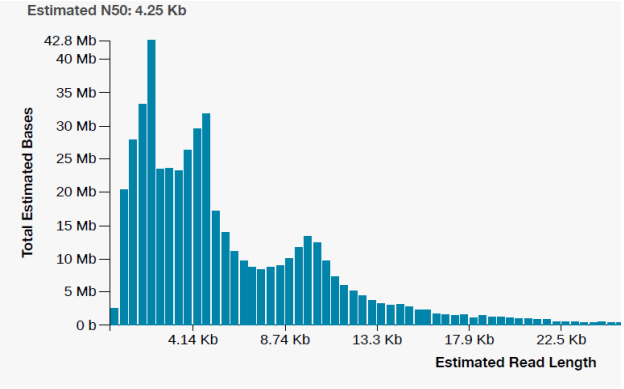

Figure B5: Read length histogram for sample 339606 P3D8

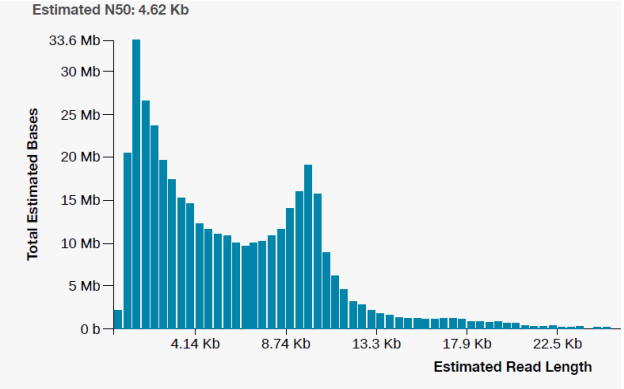

Figure B6: Read length histogram for sample 339606 P3G7

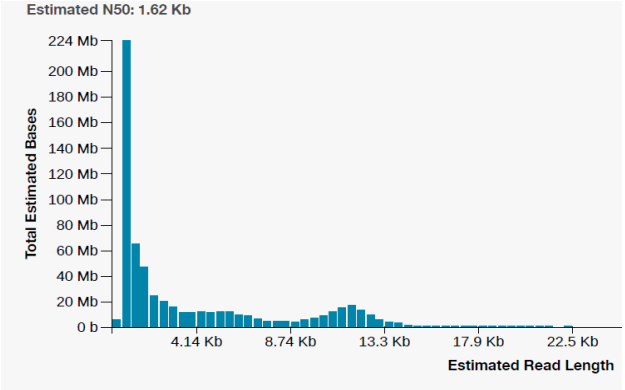

Figure B7: Read length histogram for sample 340116 P4D1

Addendum C

Table C1: Permission numbers for copyrighted images

Figure number	Reference	Permission number
1.4	Tobin & Aldrovandi, 2013	4952391206579
1.5	Murray *et al*., 2016	Not applicable Permission was obtained directly from the The Journal of Immunology